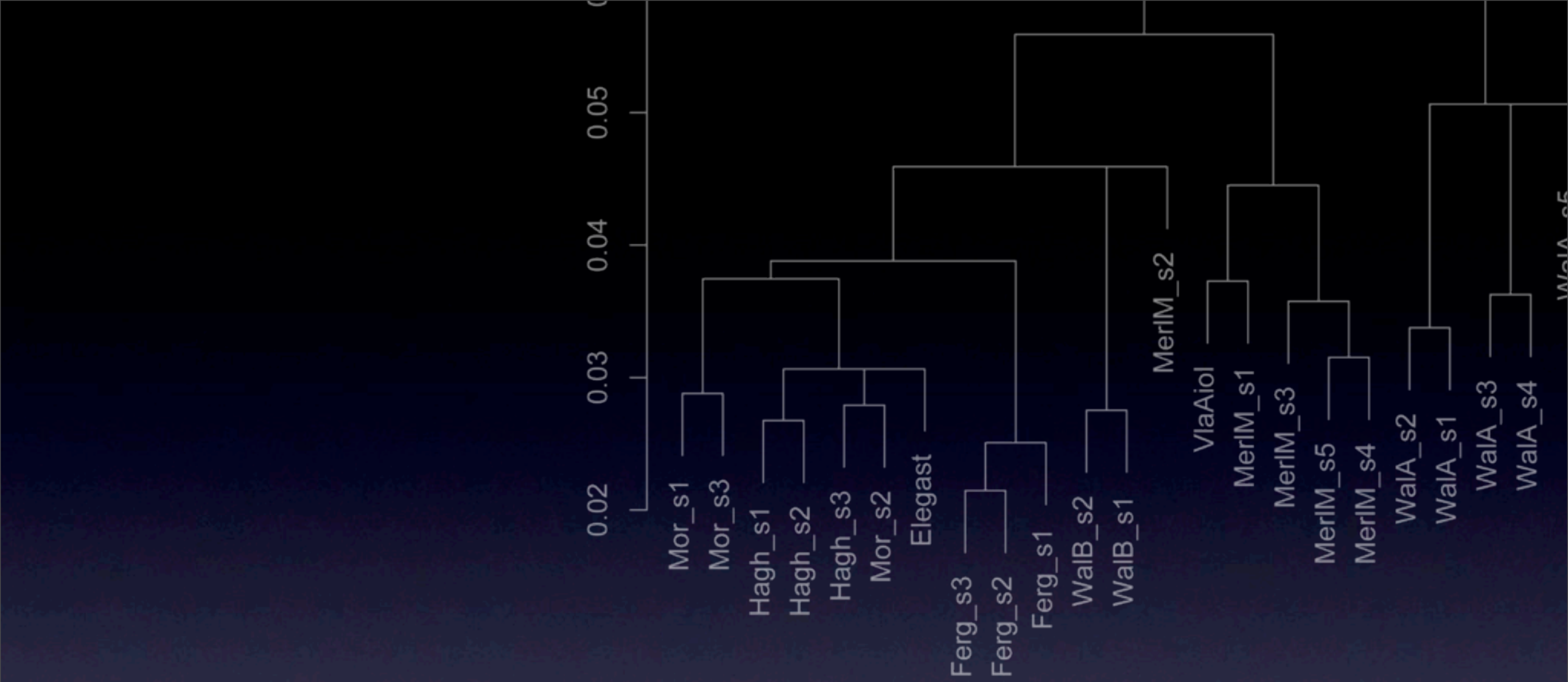


Lezen vanop afstand

Digital Humanities en de computationele analyse van middeleeuwse literatuur

www.mike-kestemont.org | www.fwo.be | www.ua.ac.be

Digital Humanities

- Digitale geesteswetenschappen
- R. Busa (1913-2011)
- IBM's *Index Thomisticus*
- Busa award door ADHO
- Mediëvist!

Gemeenplaats...

- Explosie digitale data & methodes
- “Digital humanities”
- Taal- en letterkunde
- “Traditionele” vakbeoefening
- Veel nieuw potentieel
- Maar hoe realiseren?

Distant reading

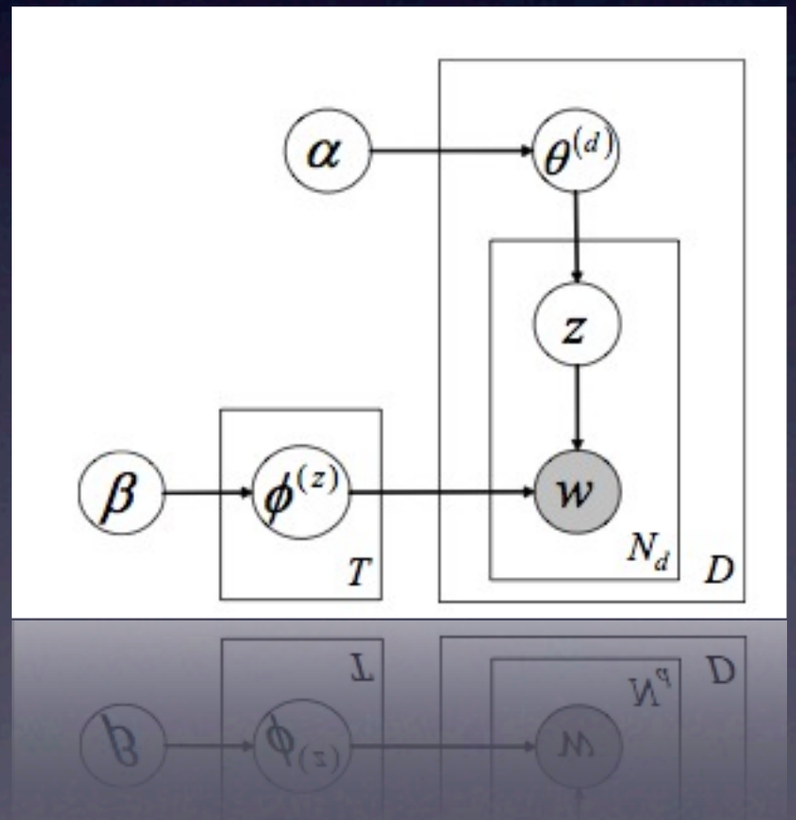
- Veelgeciteerde, 'holle' term
- F. Moretti (2000)
- *Close reading*
- Analyse van grote verzameling teksten "without a single direct textual reading"
- Uitdaging...

Topic modeling

- Recent veel aandacht in DH
- Automatische detectie van topics in groot corpus documenten
- a synecdoche of digital humanities. It is distant reading in the most pure sense: focused on corpora and not individual texts, treating the works themselves as unceremonious 'buckets of words', and providing seductive but obscure results in the forms of easily interpreted "topics" (Meeks & Weingart 2012)

Latent Dirichlet Allocation

- Dimensie-reductie (multivariate statistiek)
- *Latent Semantic Analysis, Non-negative Matrix Factorization, ...*
- Distributionele semantiek
- Monte Carlo-methode (via Gibbs sampler)
- Niet geometrisch, maar probabilistisch



Topics?

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

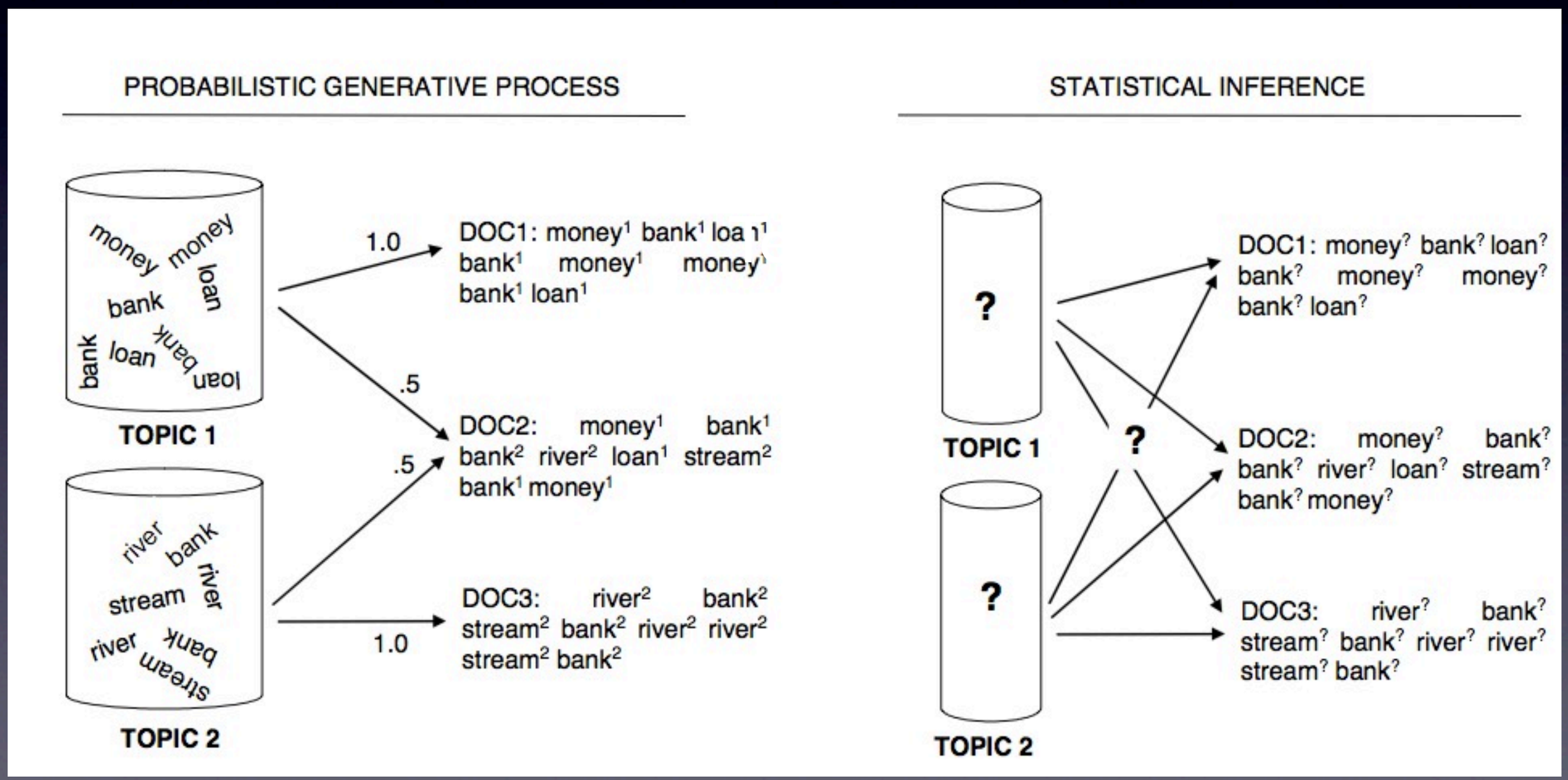
word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

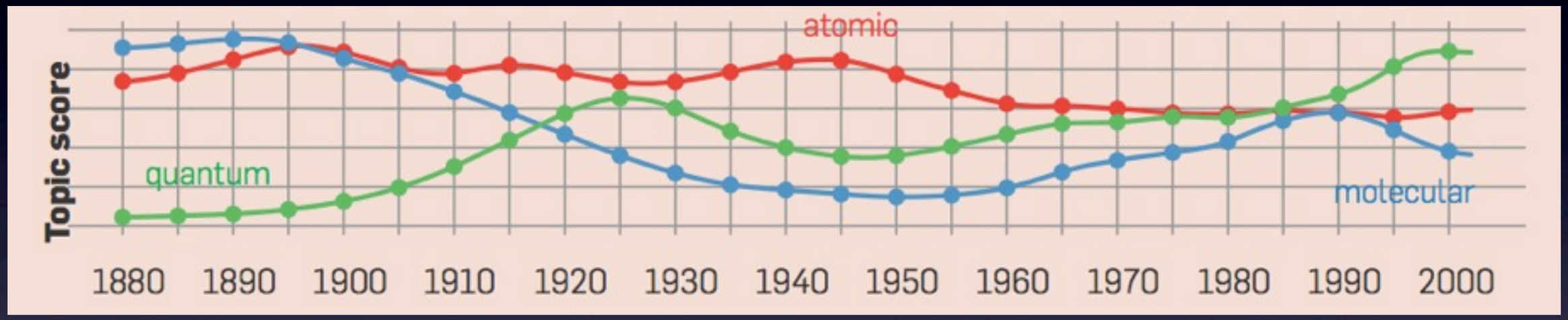
word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

Grabbelton

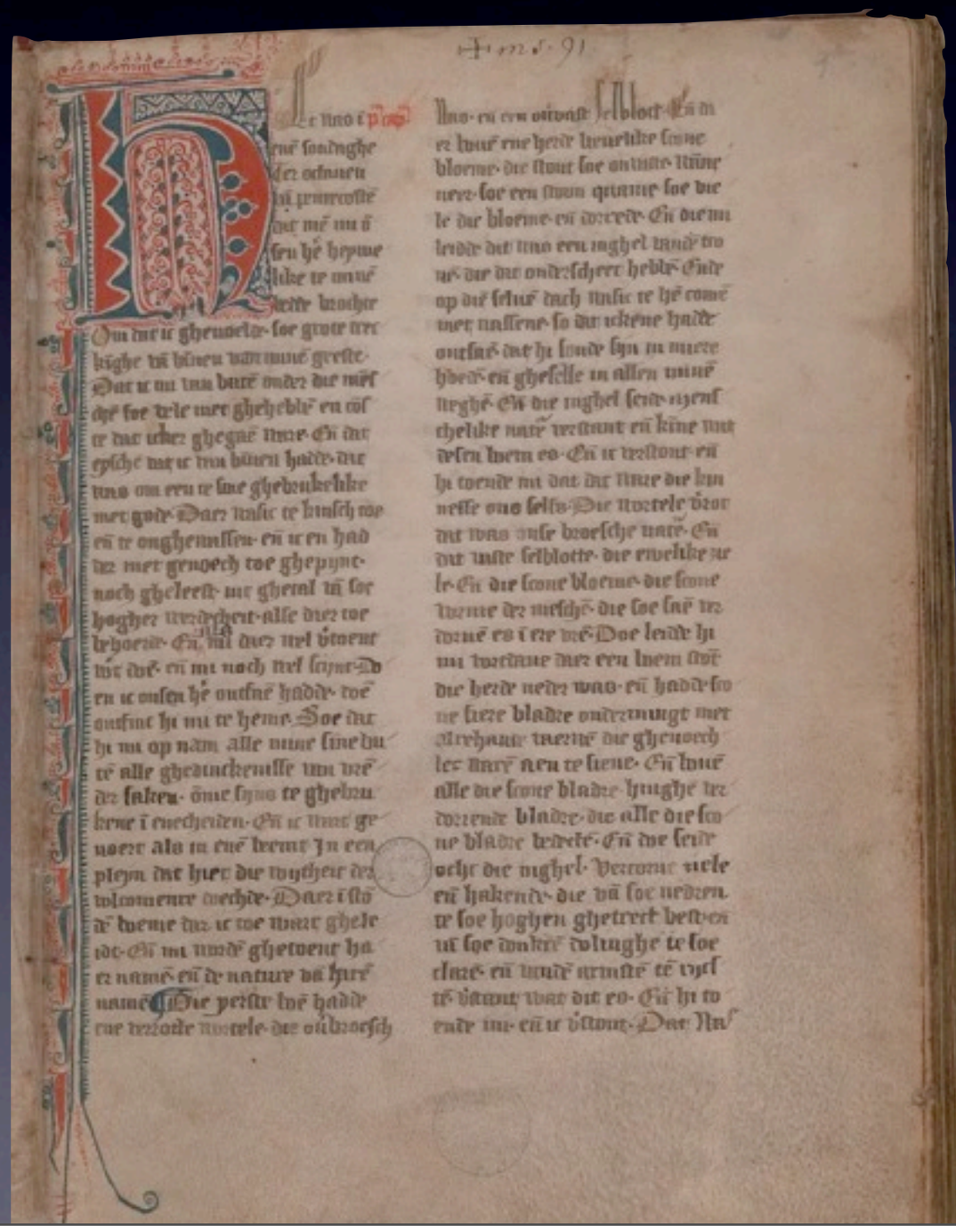

PROBABILISTIC GENERATIVE PROCESS
money money loan bank bank loan bank loan
TOPIC 1
river bank river stream river bank stream
TOPIC 2
1.0
.5
.5
1.0
DOC1: money¹ bank¹ loan¹ bank¹ money¹ money¹ bank¹ loan¹
DOC2: money¹ bank¹ bank² river² loan¹ stream² bank¹ money¹
DOC3: river² bank² stream² bank² river² river² stream² bank²
STATISTICAL INFERENCE
?
TOPIC 1
?
TOPIC 2
?
DOC1: money? bank? loan? bank? money? money? bank? loan?
DOC2: money? bank? bank? river? loan? stream? bank? money?
DOC3: river? bank? stream? bank? river? river? stream? bank?

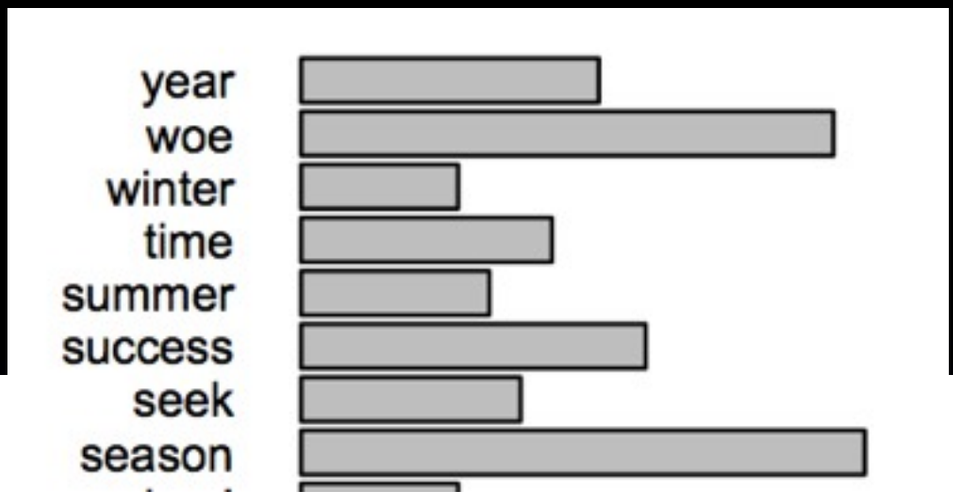
Thematische evoluties



Historische letterkunde?

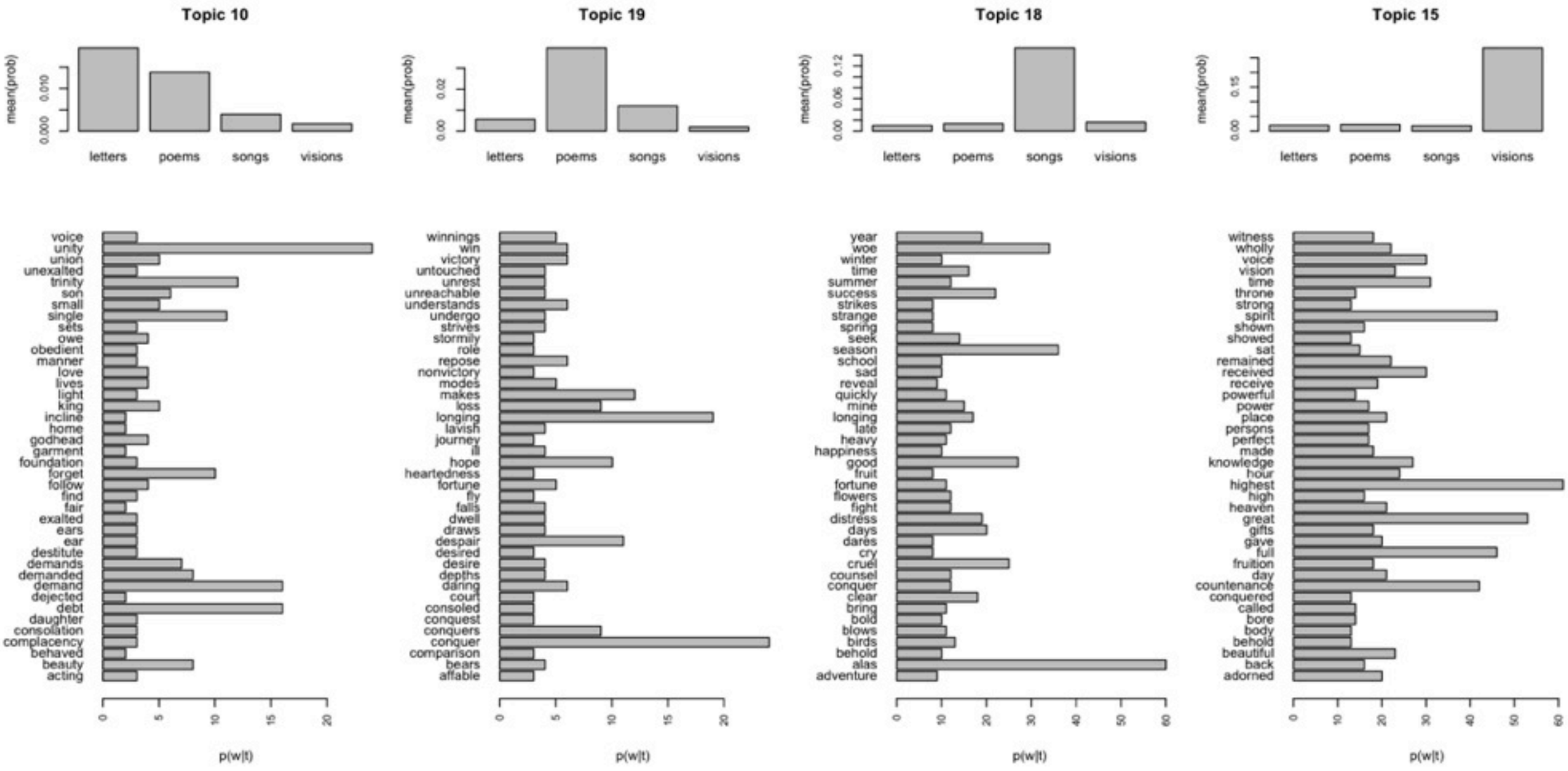
- Hadewijch (13e E)
- Brabantse mystica
- Middelnederlands
- 4 genres:
  - visioenen
  - brieven
  - gedichten
  - liederen
- Vert. C. Hart (1980)
- Traditioneel *close reading*

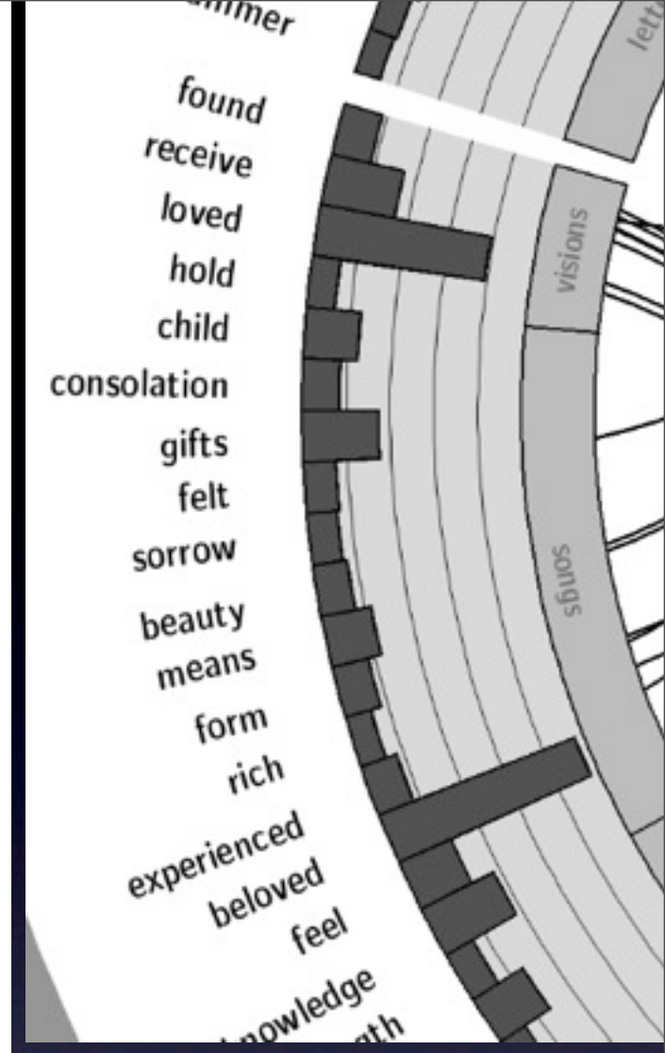




Topics in Hadewijch's Oeuvre

English translation by C. Hart (1980)





Circos

Circulaire visualisatie van genomische data

“Where’s the beef?”

- + Visuele DH (“mooie plaatjes”): bevrijdend?
- + Innovatie als drijfveer: vruchtbare confrontatie met exacte wetenschappen?
- - Wat levert het op? Nieuwigheidswaarde?
- - Verwijt: Oppervlakkigheid als Achilleshiel

Stylometrie

- Kwantitatieve studie van schrijfstijl
- Stijl <=> meta-data
  - Auteurschap (*authorship attribution*)
  - Datering (*stylochronometry*)
  - Tekstsoort (*genre studies*)
  - ...

The Evolution of Stylometry in Humanities Scholarship

DAVID I. HOLMES
The College of New Jersey, USA

Abstract

This paper traces the historical development of the use of statistical methods in the analysis of literary style. Commencing with stylometry's early origins, the paper looks at both successful and unsuccessful applications, and at the internal struggles as statisticians search for a proven methodology. The growing power of the computer and the ready availability of machine-readable texts are transforming modern stylometry, which has now attracted the attention of the media. Stylometry's interaction with more traditional literary scholarship is also discussed.

1. Introduction

The successful uncovering of Joe Klein as author of *Primary Colors* by Don Foster, the coverage by CBS TV of the UCLA conference in 1996 discussing the attribution of *A Funeral Elegy*, Robert Matthews' thoughtful articles in *Newsweek* and *The Sunday Telegraph* (see, for example, Matthews, 1994), and the request by the BBC for the Bristol Stylometry Research Unit to investigate the authorship of the 'Cassandra' letters in *Tribune* followed by footage of Richard Forsyth and myself on their *Newsnight* programme are examples of how stylometry—the statistical analysis of literary style—has begun to intrigue the media. Indeed, the age of 'pop' stylometry could now be with us.

Not all such publicity has been necessarily good, however. Channel 4's *Street-legal* programme unmasked flaws in the reliability of Andrew Morton's so-called 'cusum' technique, which had been used in UK courts in a number of high-profile criminal cases, and the Shakespeare scholar Stanley Wells recently was moved to write in the *Times Literary Supplement*:

> 'Most people who make a profession of the study of literature do so because they have an artistic rather than a scientific bent. They are, to put it simply, better at English than at Maths. But the investigation of authorship of disputed and apocryphal works relies increasingly on statistical tests that take their exponents into the realms of higher mathematics. Their works are packed with tables, charts and statistical analyses which many of those interested in the works investigated have neither the inclination nor even the intellectual ability to understand, let alone to assess.' (Wells, 1996).

Correspondence: David I. Holmes, Department of Mathematics and Statistics, The College of New Jersey, PO Box 7718, Ewing, NJ 08628-0718, USA.

Literary and Linguistic Computing, Vol. 13, No. 3, 1998

What, then, does stylometry seek to do? It certainly does not seek to overturn traditional scholarship by literary experts and historians, rather it seeks to complement their work by providing an alternative means of investigating works of doubtful provenance. At its heart lies an assumption that authors have an unconscious aspect to their style, an aspect which cannot consciously be manipulated but which possesses features which are quantifiable and which may be distinctive.

Bailey (1979) lists the general properties which quantifiable features of a text should possess: 'They should be salient, structural, frequent and easily quantifiable, and relatively immune from conscious control.' By measuring and counting these features, stylometrists hope to uncover the 'characteristics' of an author.

The two primary applications of stylometry are attributional studies and chronological problems, and Laan (1995) has provided a sound exposition of the rationale behind such applications. Laan points out that these applications seem to be based on seemingly contradictory premises, with attributional studies claiming that the unconscious aspect of an authors style remains fixed whilst chronological studies claim that stylistic features develop rectilinearly during the course of an author's life. These claims may not be incompatible; the choice of features is the overriding concern. With attributional problems, it is certainly wise to work within the same genre when faced with a list of candidate authors for a disputed work and to work within as close a time period as possible when selecting appropriate 'control' authors. Genre effects generally will supersede authorial features in the discrimination process.

The major problem inhibiting stylometry's acceptance within humanities scholarship is that, as yet, there is no consensus as to correct methodology or technique. Rudman (1997) provides a perceptive analysis of this problem and illustrates well how, for every method that 'works', there soon appear counter-arguments pointing out crucial flaws. A methodology successful for one attributional problem does not necessarily 'work' for another. Practitioners search for the 'holy grail' of stylometry, a technique beyond reproach which may be applied successfully to all genres, languages, and eras.

The historical development of stylometry is reflected in the choice of quantifiable features used as authorial discriminators. Lexical features have predominated, yet this decade has seen the application of syntactic and semantic features to attributional problems, allied to the enormous growth in computing power and the

Auteursherkenning

- Populairste toepassing
- Stylome Hypothesis
  - Unieke vingerafdruk
  - Kwantitatief meten

Journal of Quantitative Linguistics
2005, Vol. 12, No. 1, pp. 65 – 77
DOI: 10.1080/09296170500055350

Routledge

New Machine Learning Methods Demonstrate the Existence of a Human Stylome*

Hans van Halteren[1], R. Harald Baayen[2], Fiona Tweedie[3], Marco Haverkort[4] & Anneke Neijt[5]
[1]Department of Language and Speech, Radboud University Nijmegen [2]Max-Planck-Institut für Psycholinguistik, Nijmegen [3]School of Mathematics, University of Edinburgh [4]Department of Linguistics, Radboud University Nijmegen (also Department of Linguistics, Boston University) [5]Department of Dutch, Radboud University Nijmegen

ABSTRACT

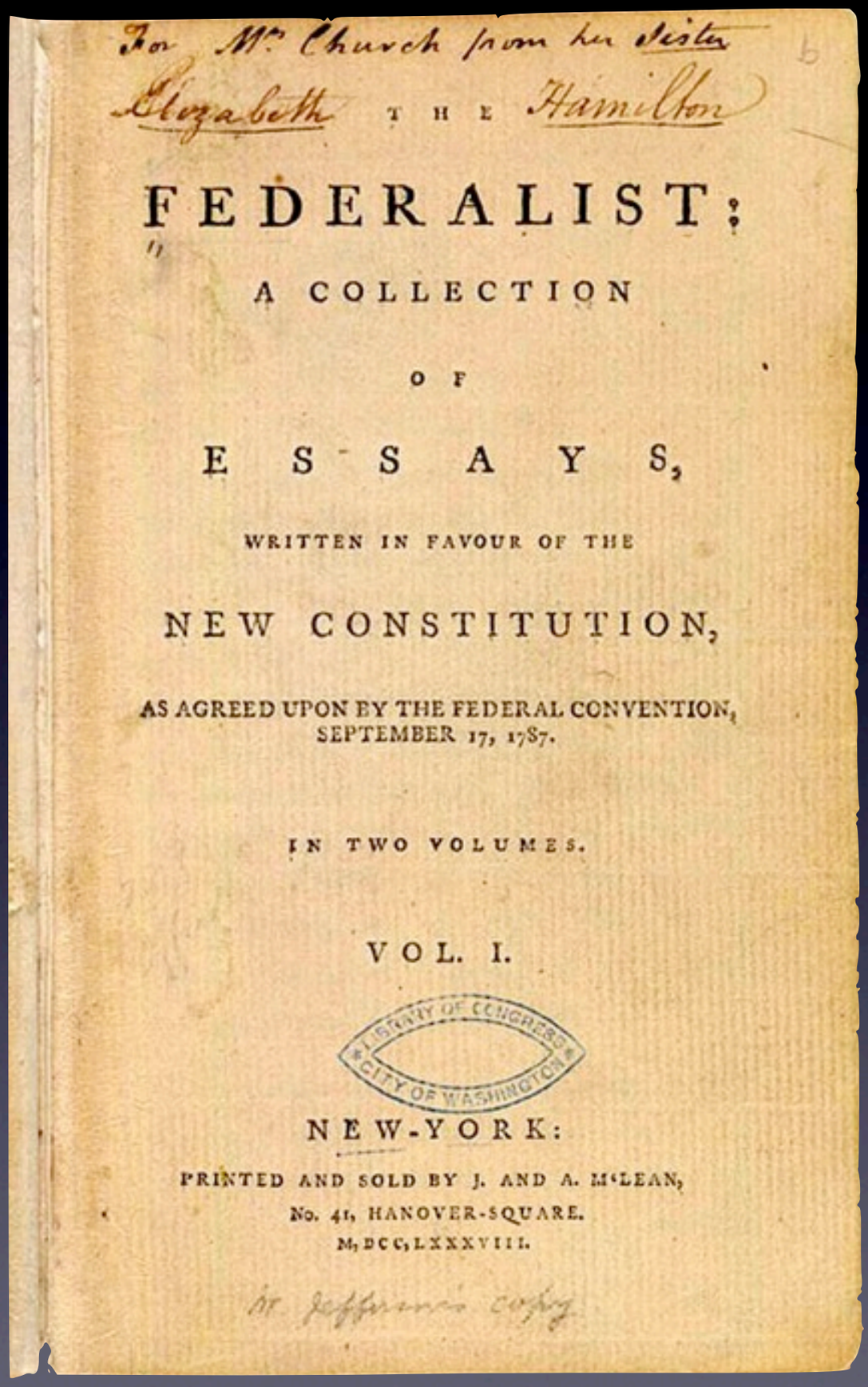
Earlier research has shown that established authors can be distinguished by measuring specific properties of their writings, their stylome as it were. Here, we examine writings of less experienced authors. We succeed in distinguishing between these authors with a very high probability, which implies that a stylome exists even in the general population. However, the number of traits needed for so successful a distinction is an order of magnitude larger than assumed so far. Furthermore, traits referring to syntactic patterns prove less distinctive than traits referring to vocabulary, but much more distinctive than expected on the basis of current generativist theories of language learning.

INTRODUCTION

We all make extensive use of a natural language to communicate with others. The ease with which we do this might give the impression that all the users of a particular natural language are using the exact same language. Now, specific languages do not come completely hard-wired in the brain, although the brain is thought to contain a set of hard-wired expectations about the structure of natural language, the so-called Universal Grammar. Whether or not there is indeed some support from structures in the brain, we always have to learn a language on the basis of

*Address correspondence to: Hans van Halteren, Department of Language and Speech, Radboud University Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen.
Tel. + 31 24361 2836. E-mail: hvh@let.kun.nl.

0929-6174/05/12010065$16.00


For Mrs Church from her sister
Elizabeth THE Hamilton

FEDERALIST:

A COLLECTION

OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

IN TWO VOLUMES.

VOL. I.

NEW-YORK:

PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M, DCC, LXXXVIII.

Mr Jefferson's copy

- Jong paradigma (1960s)
- Mosteller & Wallace (US)
- *Federalist papers* (1780s)
- Twee innovaties:
  - Kwantitatieve aanpak
  - "Functoren"

Traditioneel

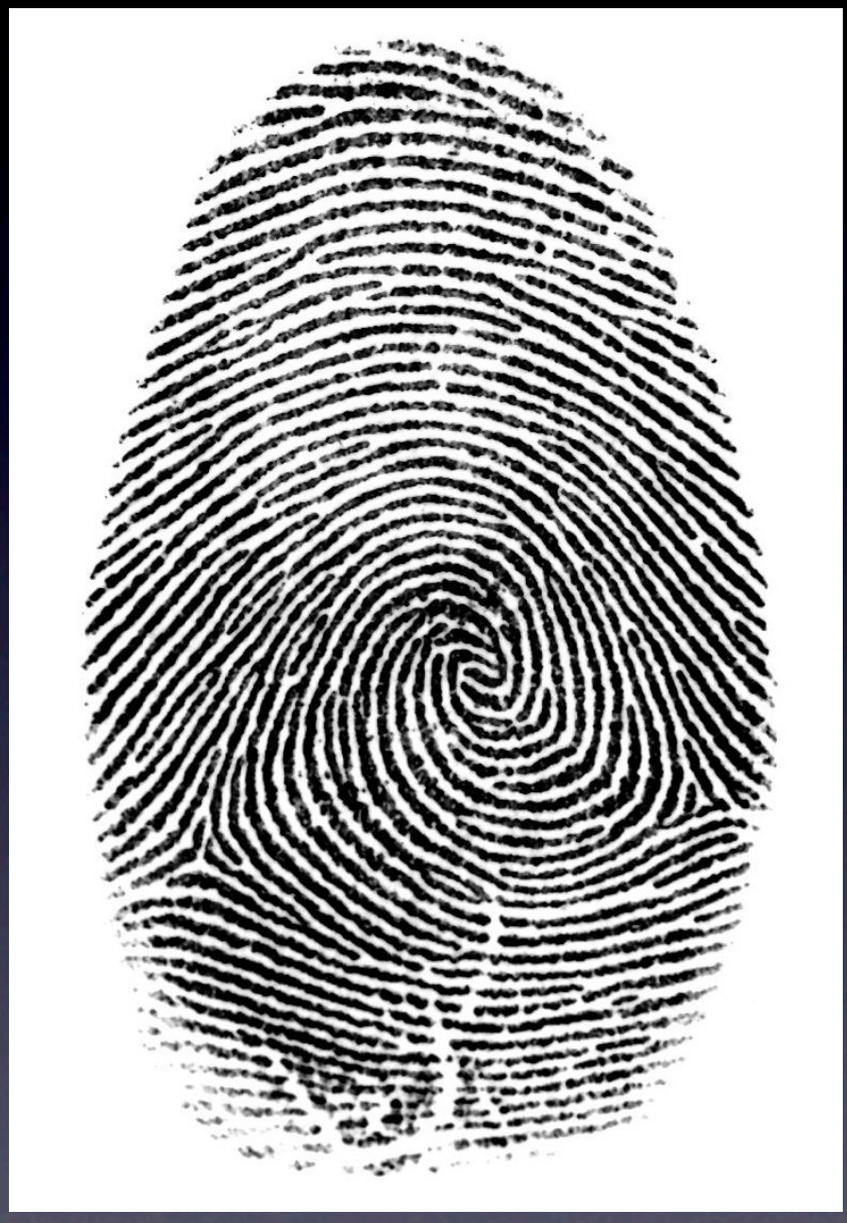
- Natte vinger...
- Opvallende kenmerken
  - bv. zeldzaam werkwoord
- “Checklist”
- Maar:
  - scholen, ateliers, ...
  - tradities
  - vervalsing, imitaties ...
  - ...

Mosteller & Wallace

- Onopvallend kenmerken
- Functiewoorden of functoren
  - lidwoorden
  - voorzetsels
  - voornaamwoorden
  - [naamvallen?]

Voordelen?

Veel observaties
Alle auteurs, zelfde set
Relatief inhoudsonafhankelijk



Aantal letters *f* op volgende slide?

Finished files are the result of years of scientific study combined with the experience of many years.

Hoeveel?

Verwerken wij functoren
'onbewust'?

Finished files are the result
of years of scientific study
combined with the experience
of many years.

Welke tekst staat op de volgende slide?

PARIS IN THE THE SPRING

En?

Moeilijk fouten detecteren...

Memory & Cognition
1977, Vol. 5 (6), 636-647

Detection errors on *the* and *and*: Evidence for reading units larger than the word

ADAM DREWNOWSKI
Rockefeller University, New York, New York 10021

and

ALICE F. HEALY
Yale University, New Haven, Connecticut 06520
and Haskins Laboratories, New Haven, Connecticut 06511

In five experiments, subjects read 100-word passages and circled instances of a given target letter, letter group, or word. In each case subjects made a disproportionate number of detection errors on the common function words *the* and *and*. The predominance of errors on these two words was reduced for passages in which the words were placed in an inappropriate syntactic context and for passages in which word-group identification was disturbed by the use of mixed typecases or a list, rather than a paragraph, format. These effects for the word *and* were not found for the control word *art*. These results were taken as evidence that

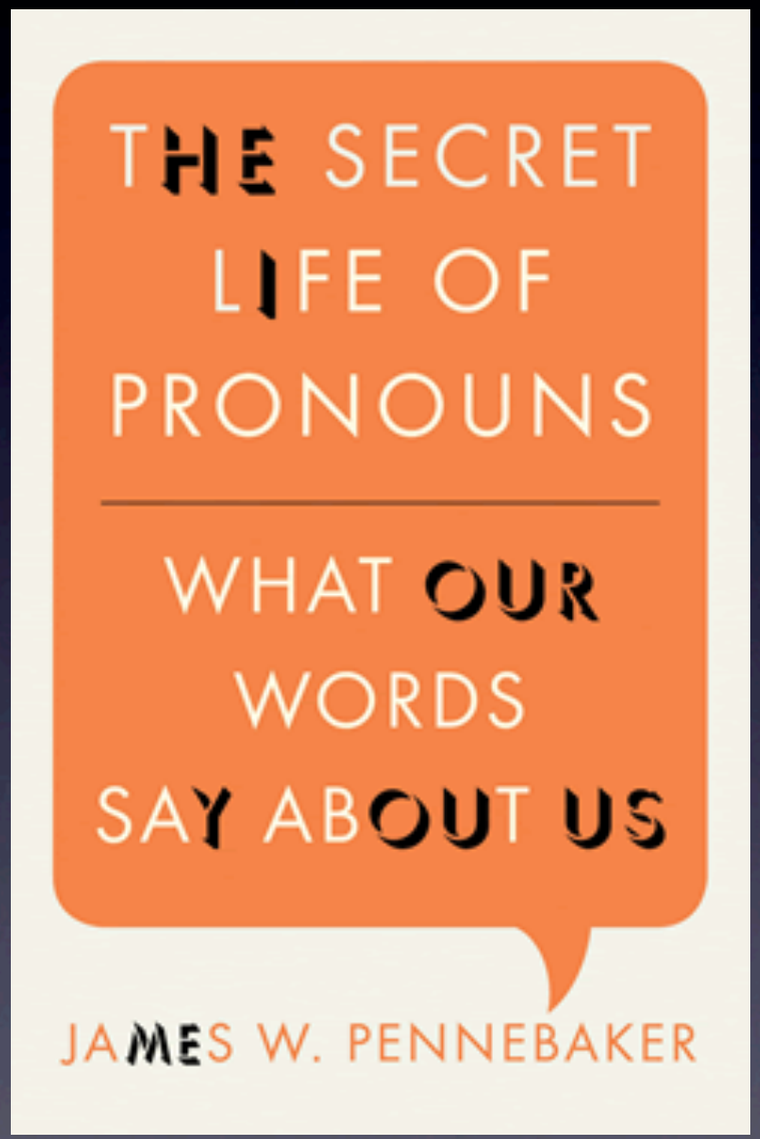
Onbelangrijk?

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

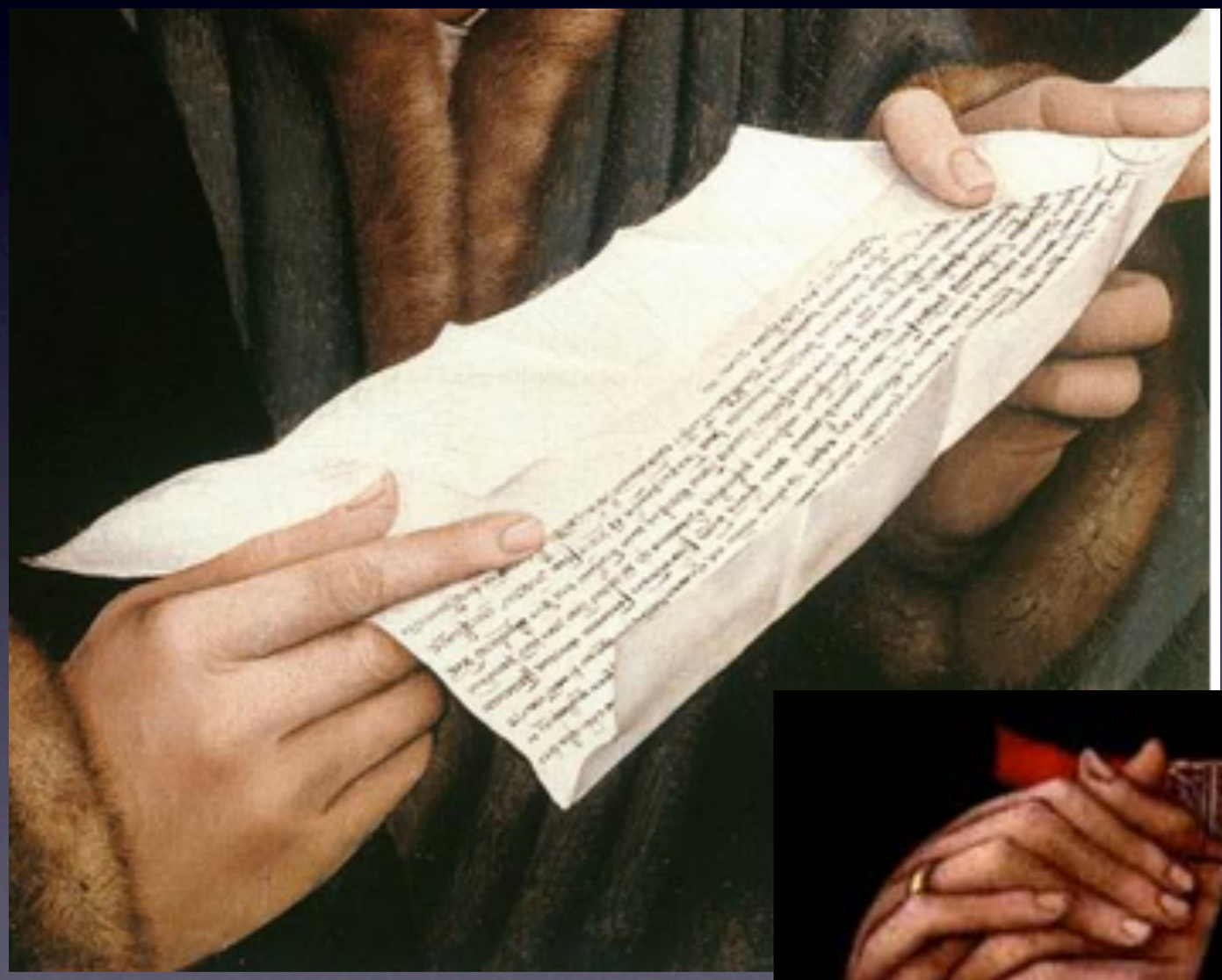
Functoren

- Populair in stylometrie
- Frequenties als input
- Pennebaker (2011)



Parallel in kunstgeschiedenis

Morelli (1816–1891)



Middelnederlands

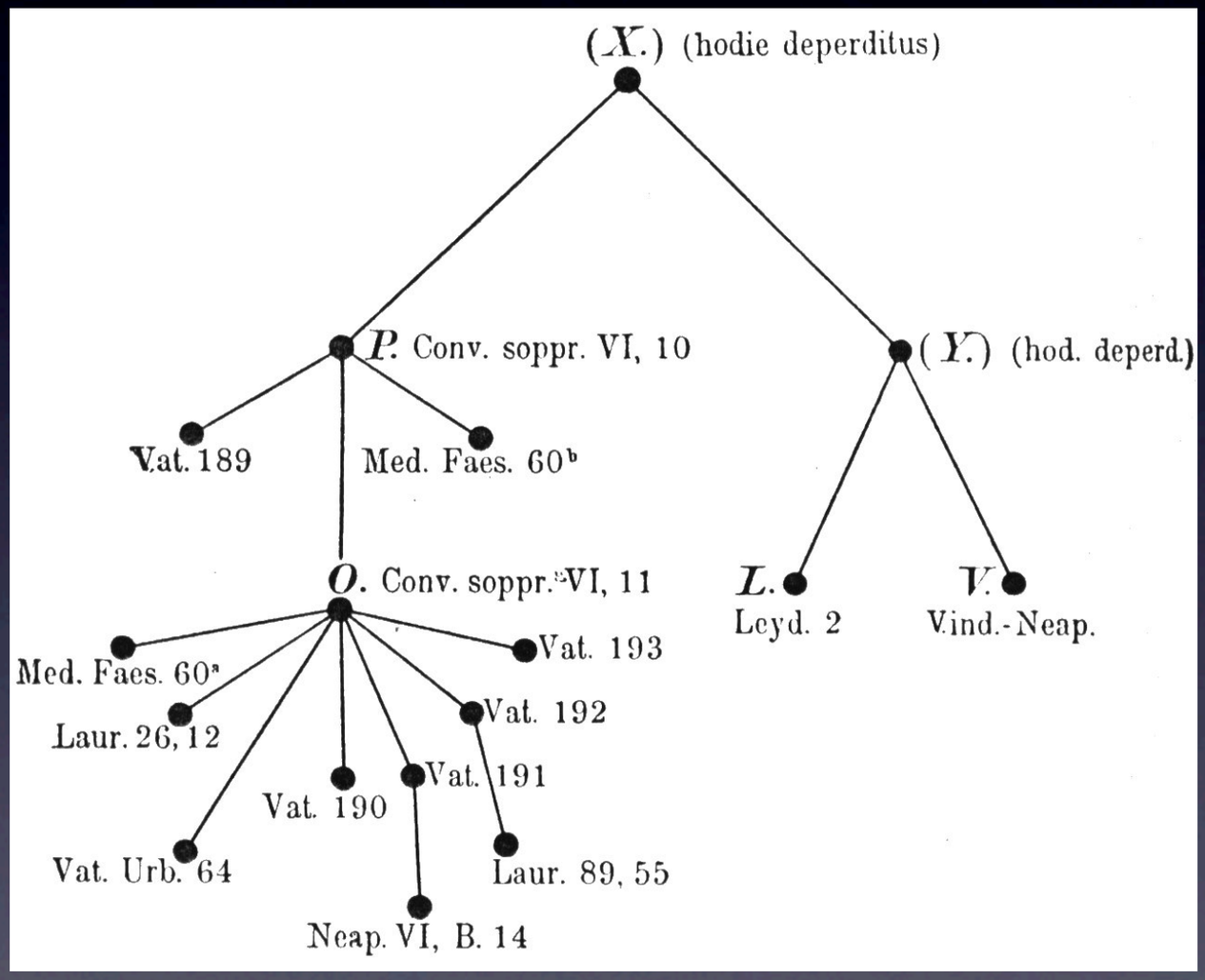
- Germaanse volkstaal
- Lage landen
- Middeleeuwen
- Ca. 1200-1500
- Literatuur

Handgeschreven wereld

- Geen drukpers
- Manu-scripten
- Kopiisten scribenten
- Iedere kopie uniek

Variatie als regel

- Geen standaardtaal, -spelling
- Kopieën wijken af van legger:
  - Spelling
  - Locale dialecten
  - Stilistische voorkeur?
  - ...
- Tekst oorspronkelijke auteur?



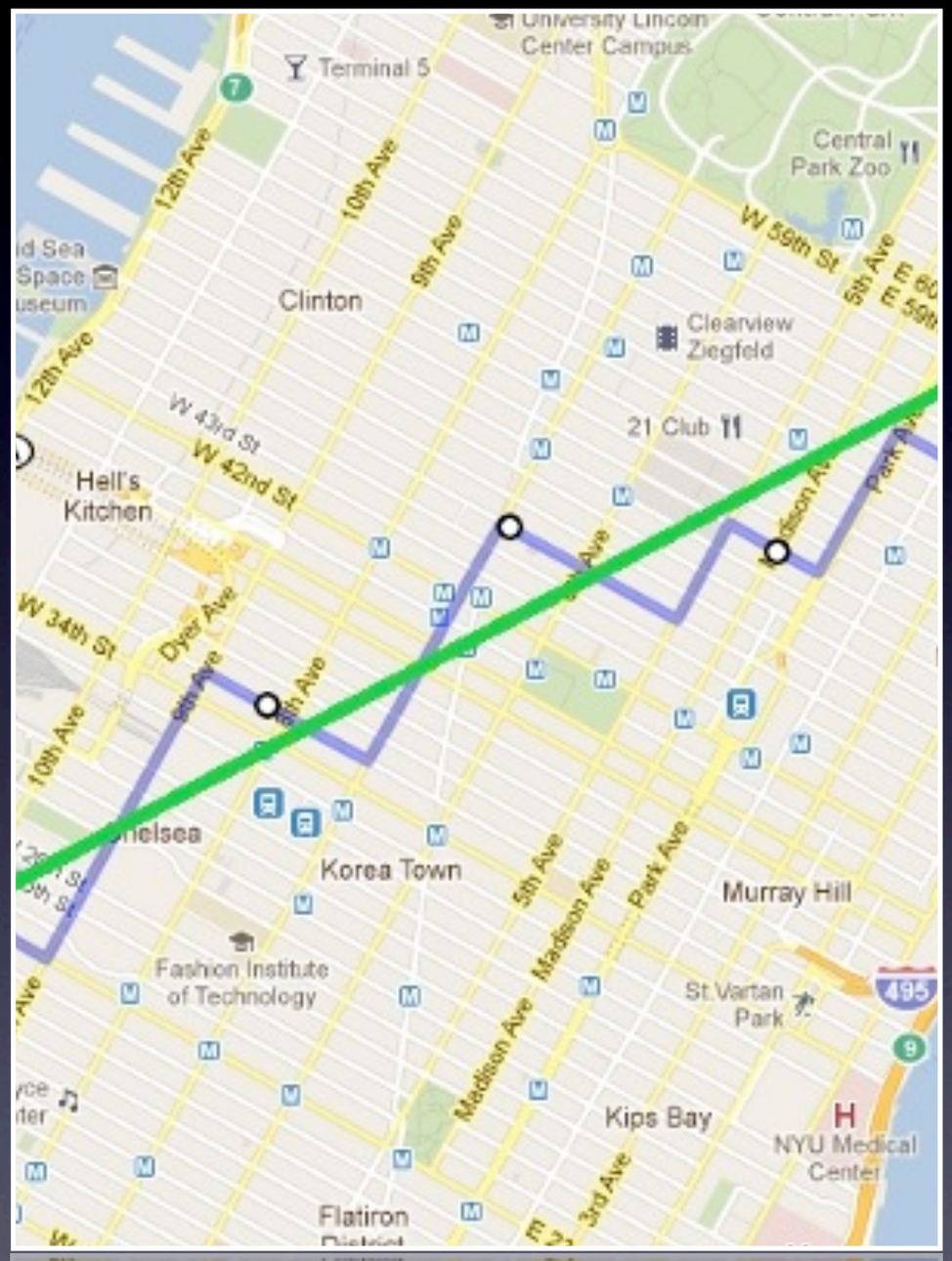
Middelnederlands

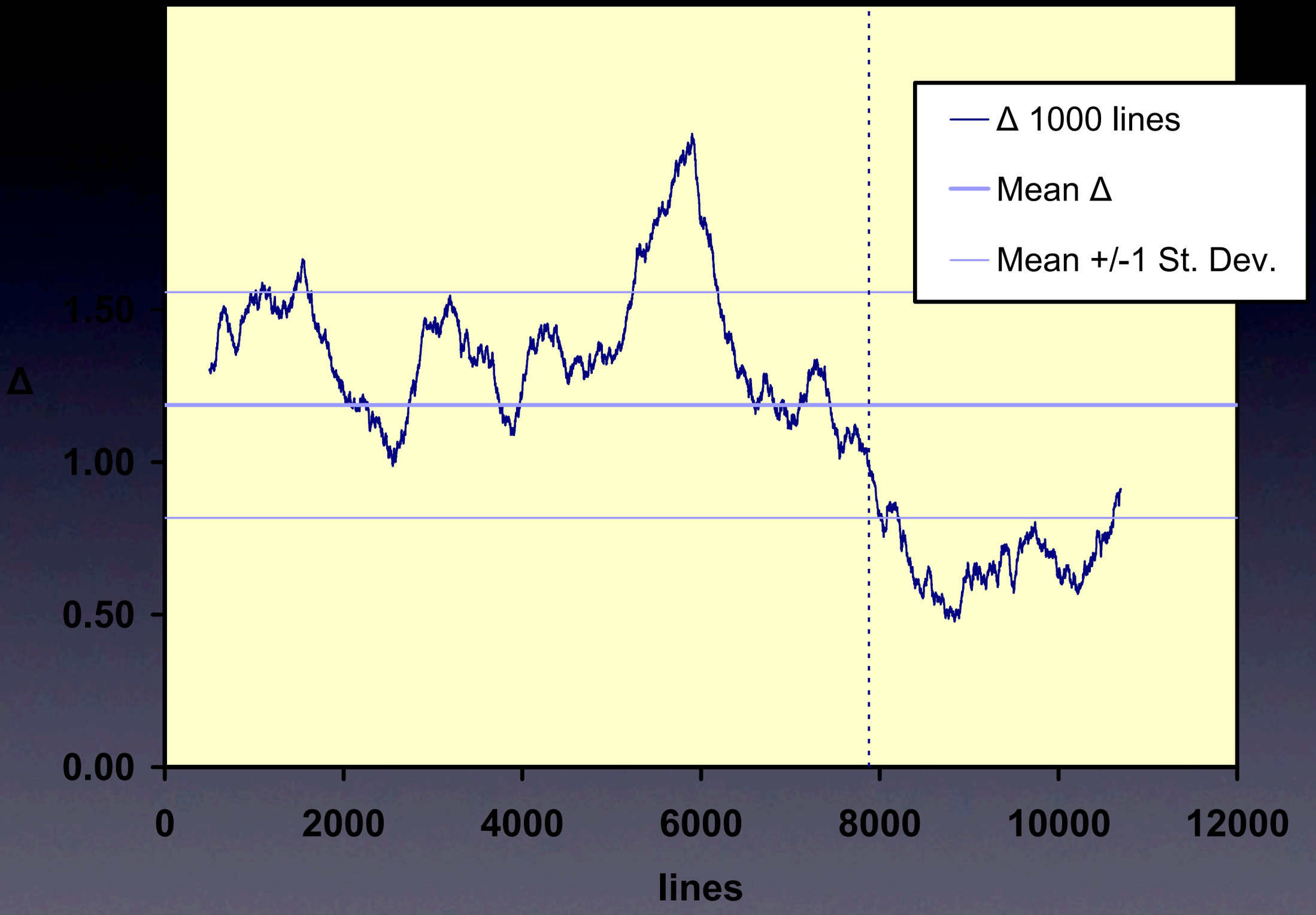
- Van Dalen-Oskam & Van Zundert 2007
- *Literary and Linguistic Computing*
- *Roman van Walewein*
- Auteursovergang
- Leiden, Ltk. 195
- Pionierswerk

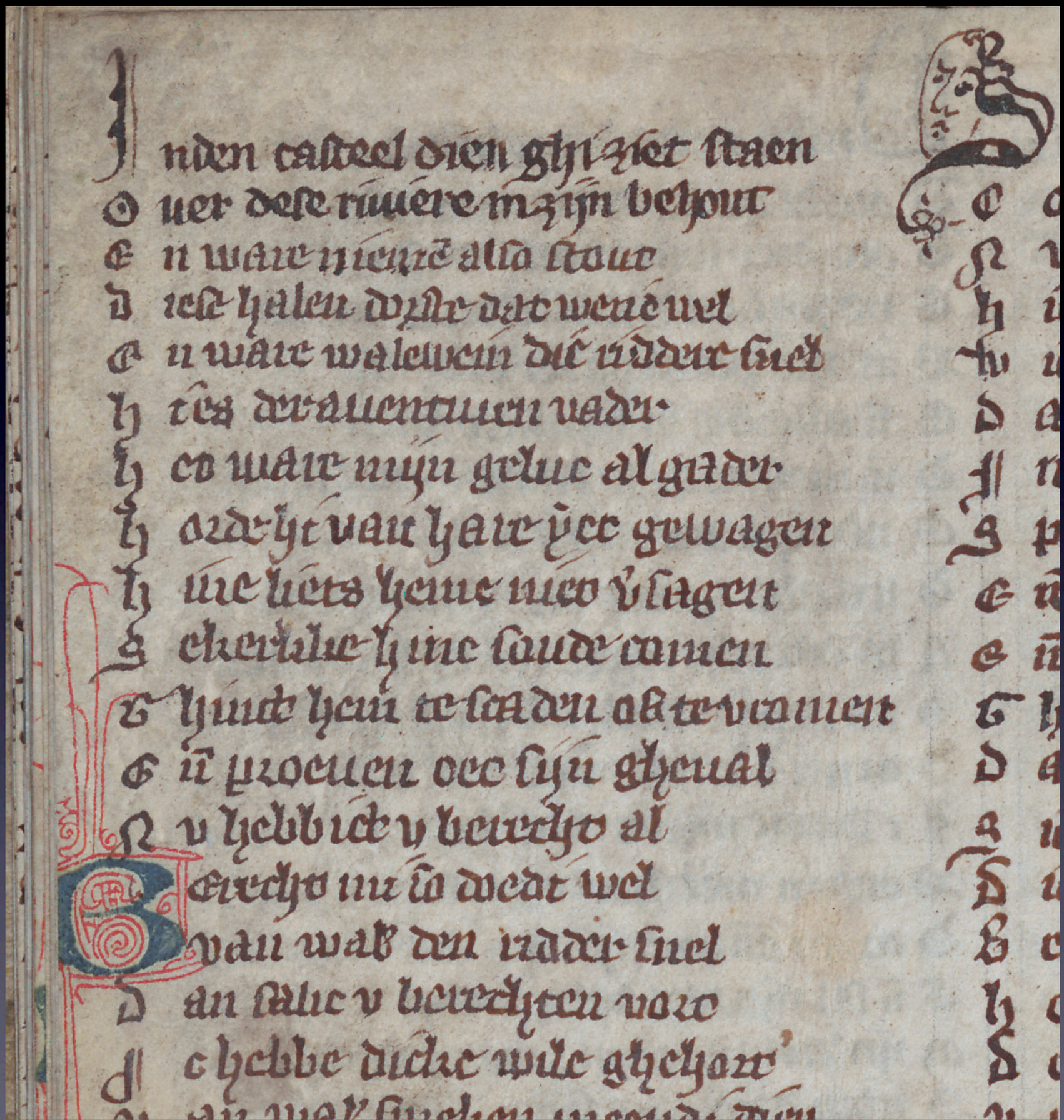
Burrows's Delta

- Stilistische afstandsberekening
- *Manhattan distance* op MWF
- *Nearest neighbor learning*
- Zowel classificatie als afstand

$$\Delta_B^{(n)}(D, D') = \sum_{i=1}^{n} \frac{1}{\sigma_i} \left| f_i(D) - f_i(D') \right|$$

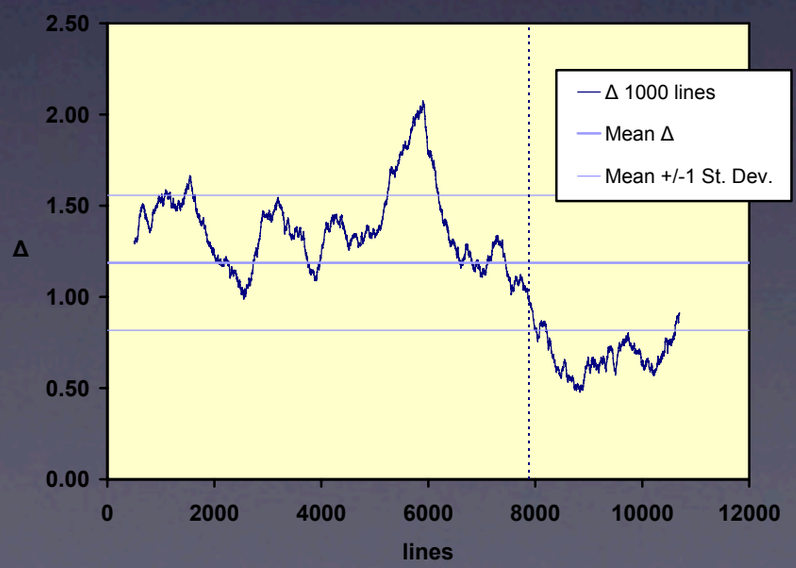



Δ 1000 lines
Mean Δ
Mean +/-1 St. Dev.
Δ
1.50
1.00
0.50
0.00
0
2000
4000
6000
8000
10000
12000
lines



Leiden

UB, Ltk. 195



Serendipiteit?

Zochten auteurs en vonden kopiisten...

Delta for Middle Dutch—Author and Copyist Distinction in *Walewein*

Karina van Dalen-Oskam and Joris van Zundert
Huygens Instituut, The Hague, The Netherlands

Abstract

The Middle Dutch Arthurian romance *Roman van Walewein* ('Romance of Gawain') is attributed in the text itself to two authors, Penninc and Vostaert. Very little quantitative research into this dual authorship has been done. This article describes our progress in applying different non-traditional authorship attribution methods to the text of *Walewein*. After providing an introduction to the romance and an overview of earlier research, we evaluate previous statements on authorship and stylistics by applying both Yule's measure of lexical richness and Burrows's Delta. To find out whether these new methods would confirm or even enhance our present knowledge about the differences between the two authors, we applied an adapted version of John Burrows's Delta procedure. The adapted version seems to be able to distinguish the double authorship of the romance. It also helps us to confirm some and to reject other earlier statements about the position in the text where the second author started his work.

Correspondence:
Karina van Dalen-Oskam, Huygens Instituut, Postbus 90754, NL-2509 LT Den Haag, The Netherlands.
E-mail:
karina.van.dalen@huygensinstituut.knaw.nl

1 Introduction[1]

Until now, non-traditional authorship attribution techniques have hardly been applied to medieval texts. The fact that we do not and probably never will know the author of many texts, may play a role in this. There are cases, however, in which the researcher does not necessarily want to identify a specific author but wishes to distinguish one (perhaps anonymous) author from another. In these cases, it could indeed be very interesting to turn to modern authorship attribution techniques and find out of how much help they actually can be.

In our research on medieval Dutch literature, we have long been intrigued by a Middle Dutch Arthurian romance that seems to be the perfect case for an experiment like this. In this article, we first present the case and the questions we started out with (Section 2). We then summarize our first results and present the new questions to which these lead (Section 3). In Section 4, we investigate two specific questions. We then focus on the possibilities of Burrows's Delta for distinguishing the authors as well as the scribes of *Walewein* (Section 5). After presenting our results, we sum up our conclusions and describe how we plan to test our results and to refine the method we are developing.

2 The Case: *Roman van Walewein*

One of the few Middle Dutch romances that were not translated from Old French is the Arthurian romance known as *Walewein* or *Roman van Walewein* ('Romance of Walewein'). The main character of this verse text in rhyming couplets is Arthur's nephew Walewein—the 'Gawain' of English texts and the 'Gauvain' of the French. *Walewein* was probably written around the year 1260, but the only surviving complete manuscript dates from almost a century later. The manuscript was written by two scribes. The first wrote lines

Literary and Linguistic Computing, Vol. 22, No. 3, 2007.
doi:10.1093/llc/fqm012 Advance Access Published on 2 June 2007

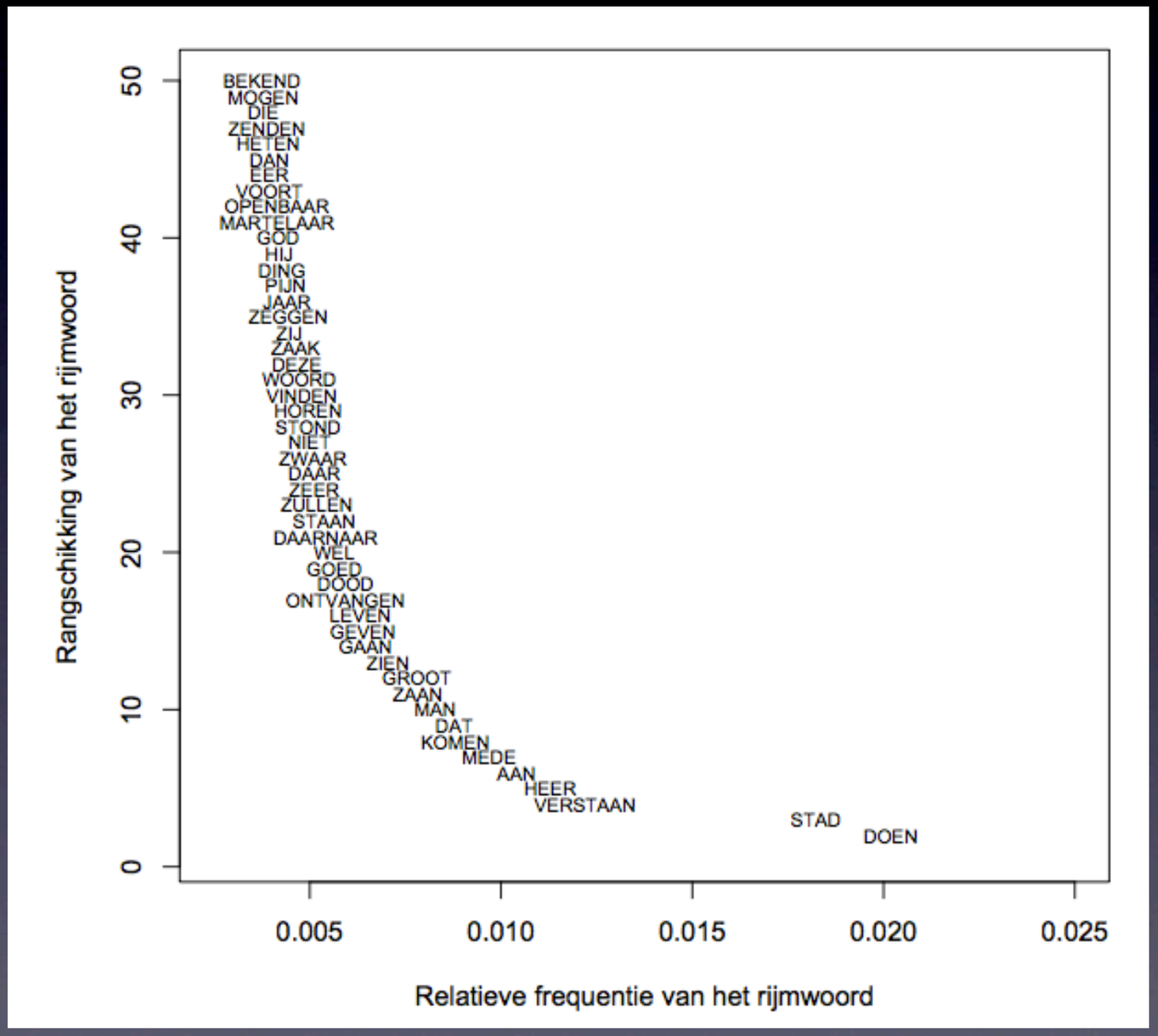
D	Ter stont ende ter seluer vren
E	Tier stont ende ter seluer vren
F	TIere stont enter seluer vren
G	Tottien stonden en ter uren
H	TEn stonden ende ter seluer vren
I	Tjerst stont ende tier veren
J	Tyer stont ende tier seluer vren
N	TJer stont tier seluer vre

Auteursherkenning vs. Scribentherkenning...

Rijmwoord

- Meeste Middelnederlandse literatuur berijmd
- Bijbel, encyclopedie, liefdesgedichten, ...
  - Willam die madock maecte / Dair hi dicke om waecte
- Skelet van tekst
- Heel moeilijk aan te passen
- “Oorspronkelijke dichter in het aangezicht staren”?

Rijmwoorden als functors?



- Eindig # combinaties
- Formules, “stoplap”
- Popliedjes: “De lucht is blauw : Ik hou van ...”
- Zipfiaanse verdeling
- Surrogaat functoren?

Lemmatiseren

- Spellingvariatie
- Lemmatiser
- Lemma-frequenties

in der ierster partien inden	EIND
mogedi van claudiusse vinden	VINDEN
hoe hi was keyser ende here groot	GROOT
dese heeft gedaen so vor sine doot	DOOD
dat nero keyser na hem blive	BLIJVEN
bi den rade van sinen wive	WIJF
ende der heren daer hi in desen	DEZE
te seer af scheen bedwongen wesen	WEZEN
octavien siere dochter man	MAN
was nero ende also dan	DAN

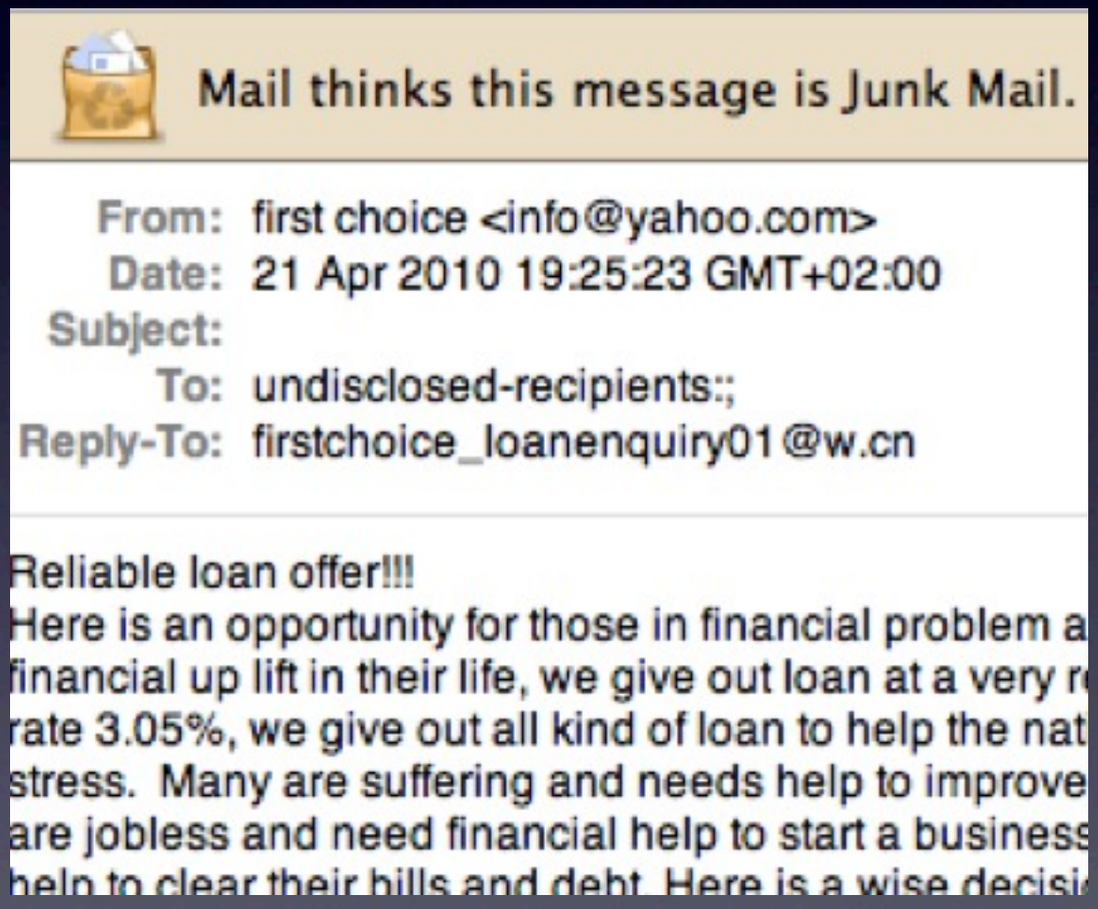
Spiegel historiael

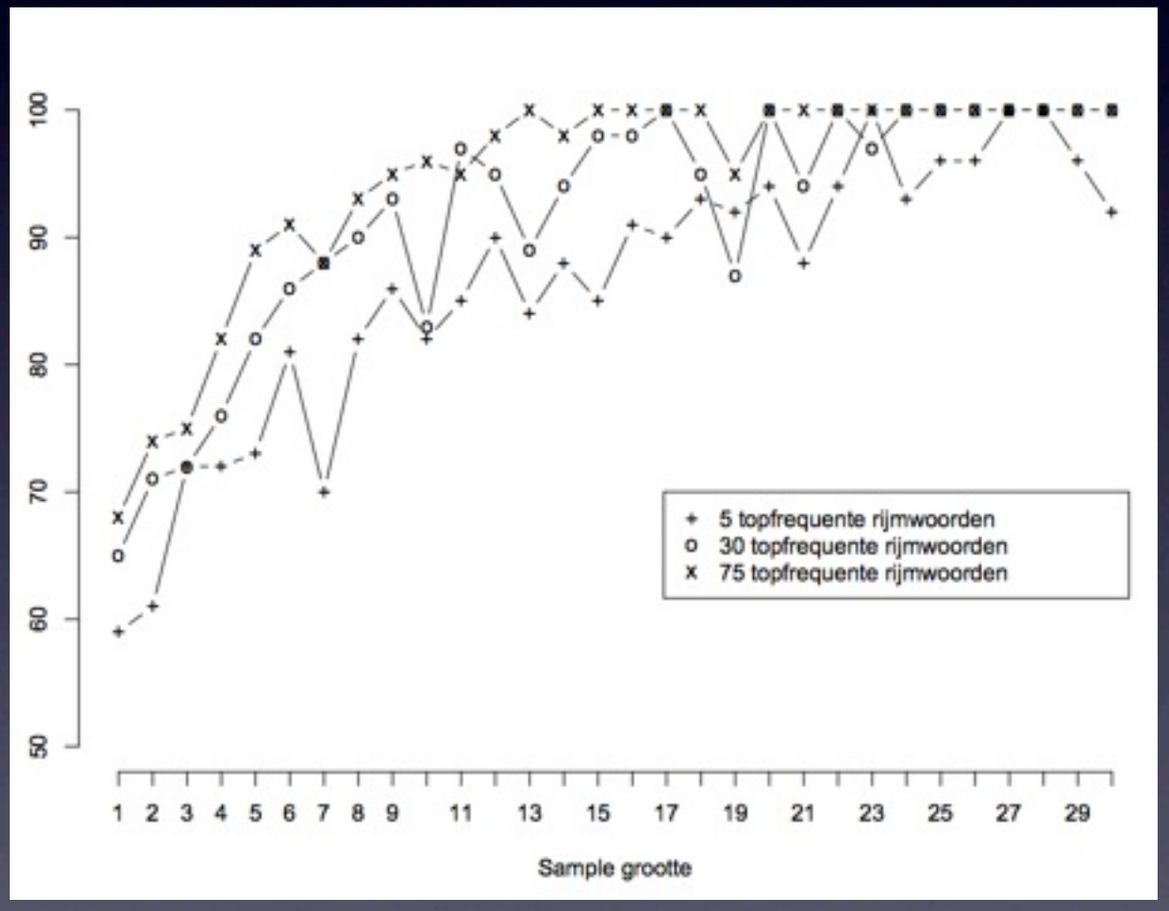
Speculum historiale (13e E)

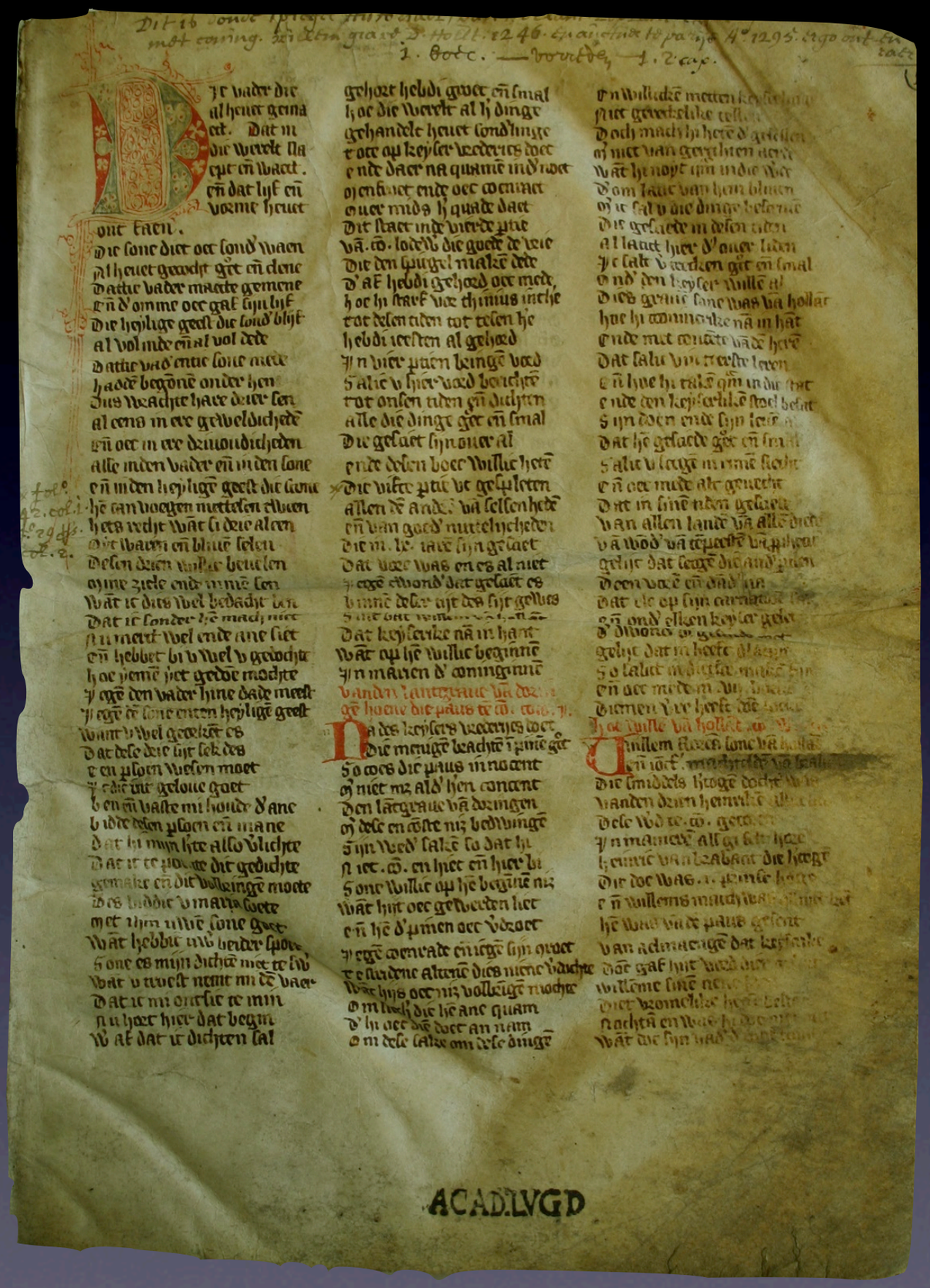
Jacob van Maerlant | Filip Utenbroeke

Tekstclassificatie

Spam filtering

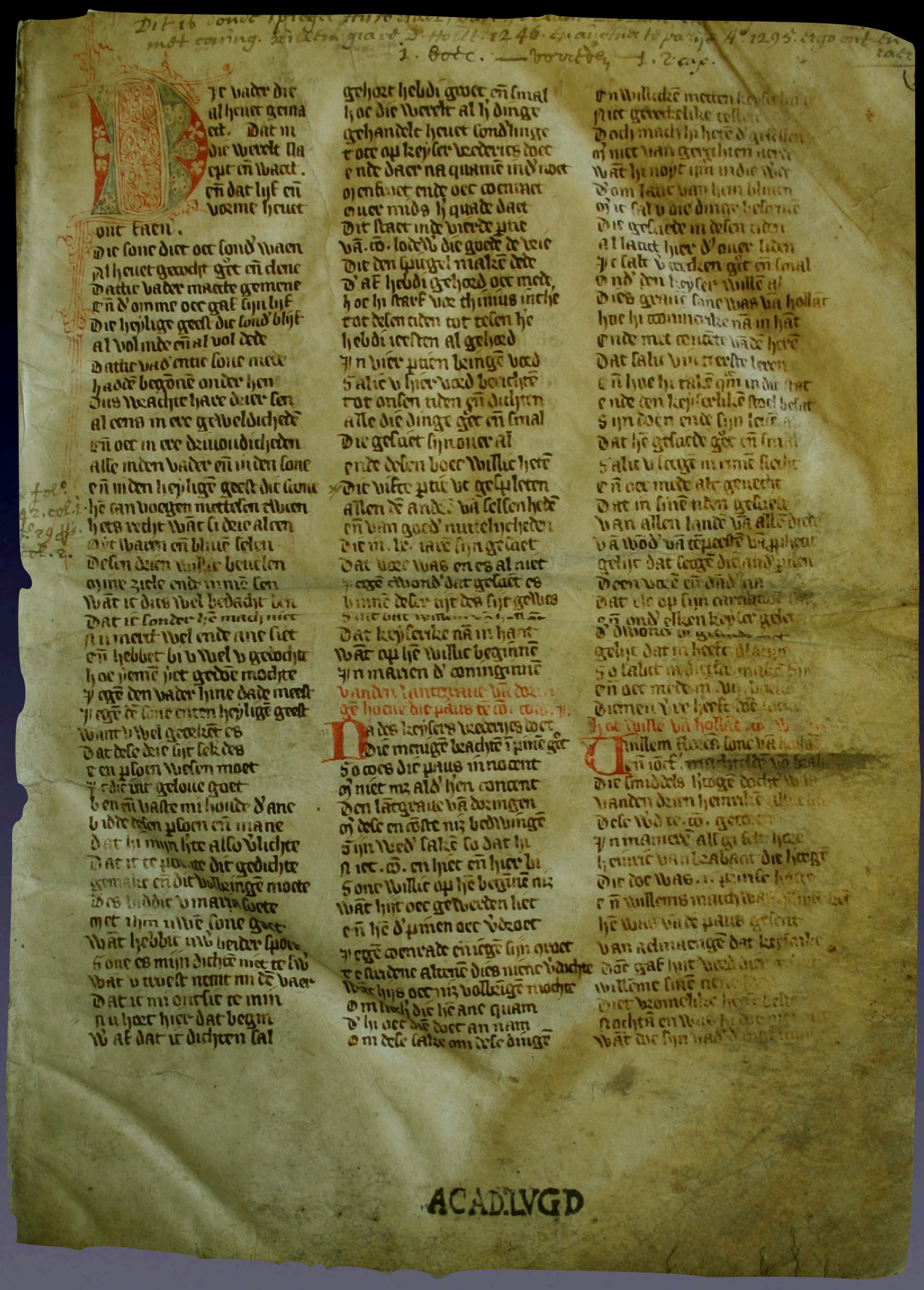






ACAD.LVGD

Laatste deel, ca. 1316: Lodewijk van Velthem Vierde en Vijfde Partie



Top 25

	Utenbroeke	Maerlant	Velthem
Utenbroeke	15	0	0
Maerlant	0	15	0
Velthem	1	0	14

Top 50

	Utenbroeke	Maerlant	Velthem
Utenbroeke	15	0	0
Maerlant	0	15	0
Velthem	0	1	14

Top 75

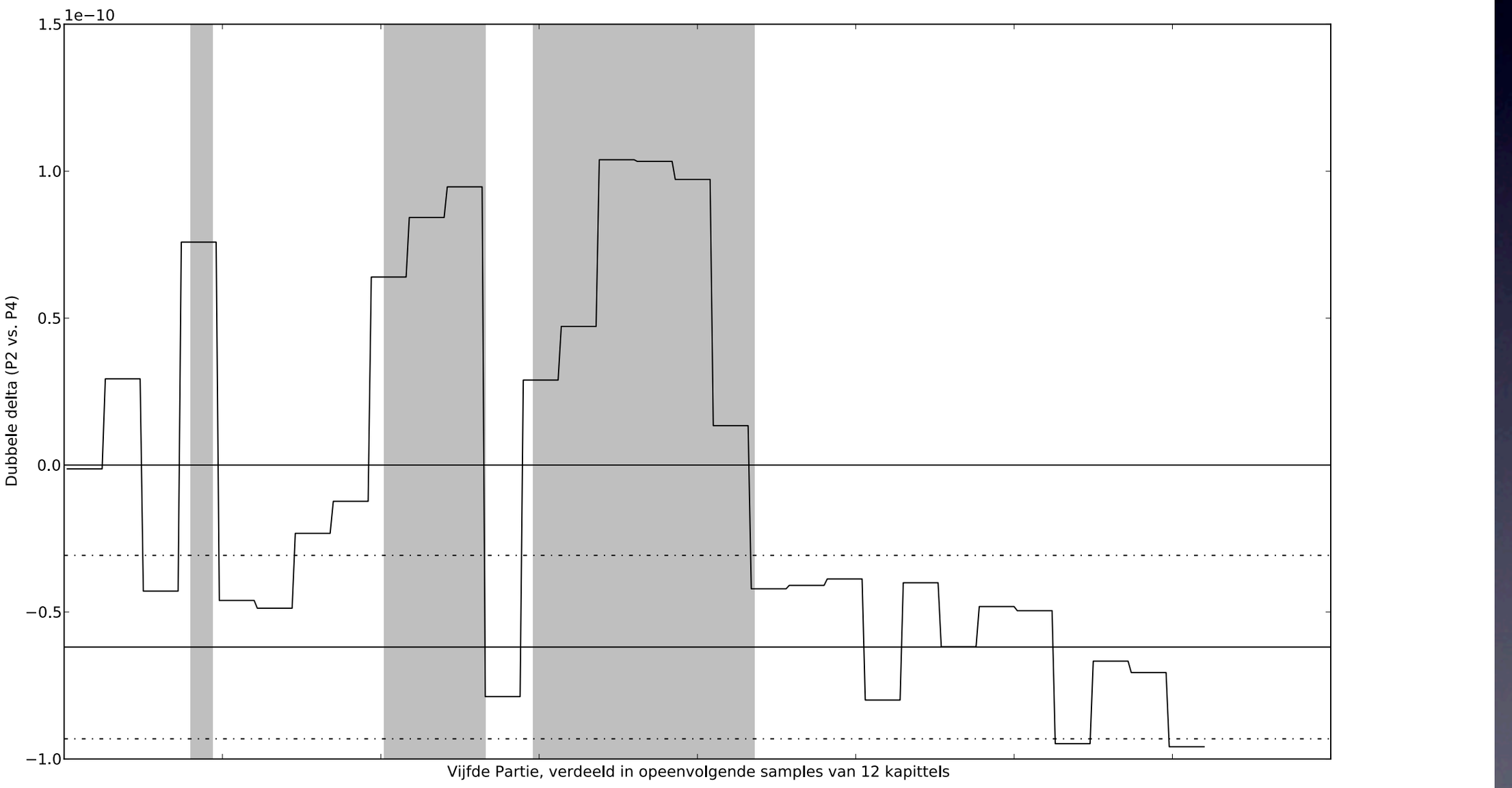
	Utenbroeke	Maerlant	Velthem
Utenbroeke	15	0	0
Maerlant	0	15	0
Velthem	0	1	14

4de boek steeds “fout” toegeschreven...

Guldensporenslag, 1302

Jan van Heelu
Slag bij Woeringen
1288

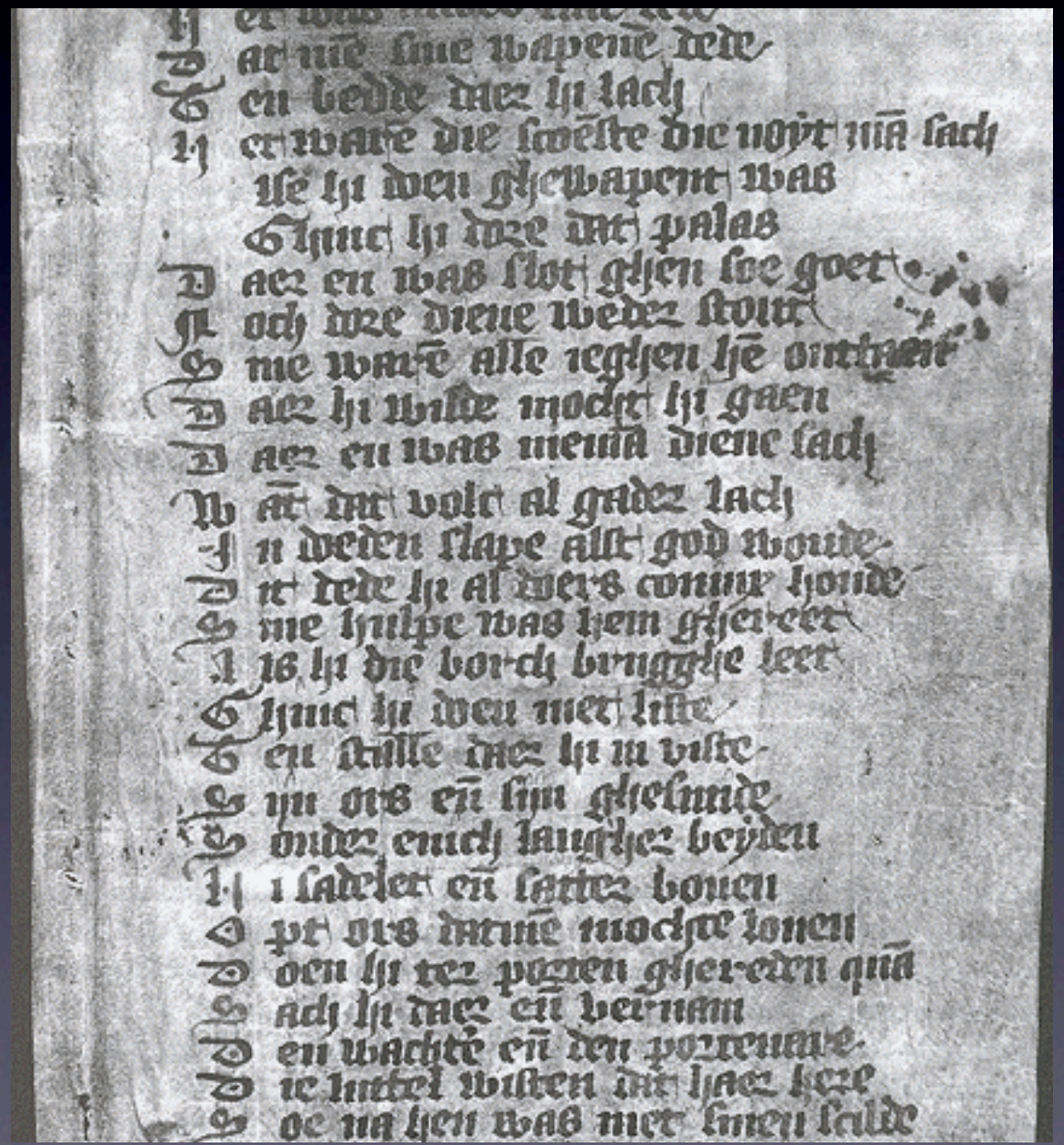
Stijlcurve Velthems Vijfde Partie...



“Ontleende” Velthem een reeds bestaand Vlaams ooggetuigeverslag? (serendipiteit...)

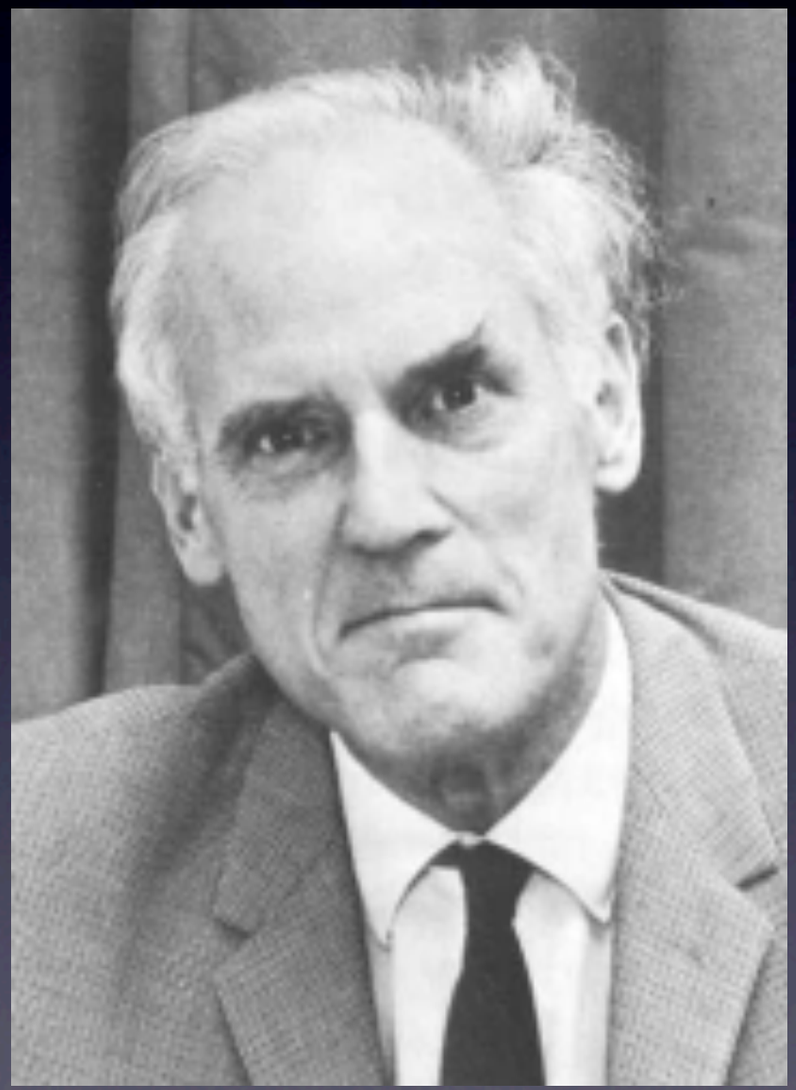
Karel ende Elegast

- Icoon Mnl'se literatuur
- 13e eeuw (?), Vlaanderen
- Anoniem, berijmd verhaal
- *Fraeye historie ende al waer*
- Nachtelijke rooftocht van Karel de Grote en Elegast



K.H. Heeroma (1909-1972)

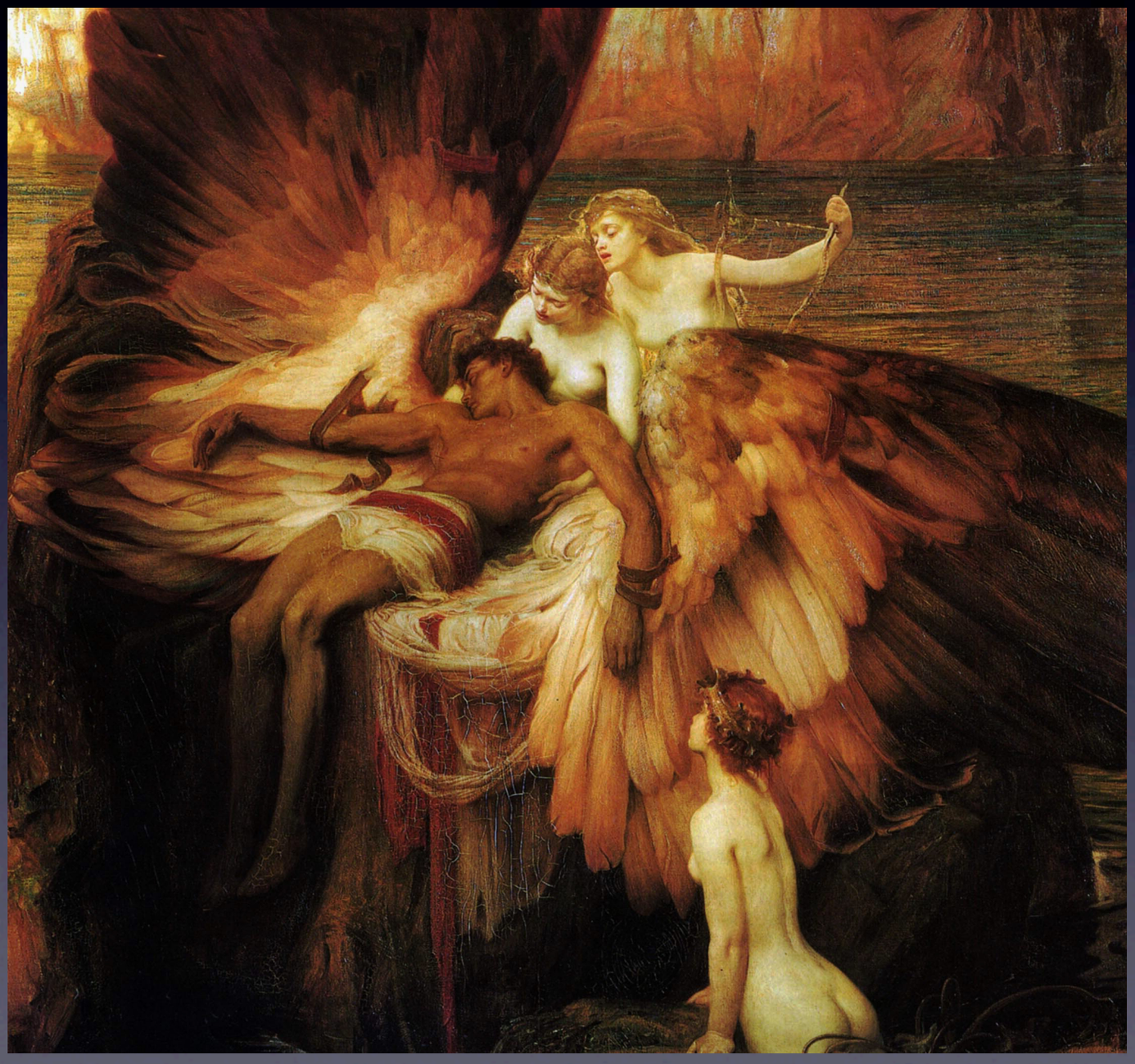
- Nederlands filoloog, zelf dichter
- "Subjectieve" stilistische analyse
- Controversieel
- "Hoorde" zelfde stem in *Elegast* en *Moriaen*
- Zelfde auteur?



Moriaen

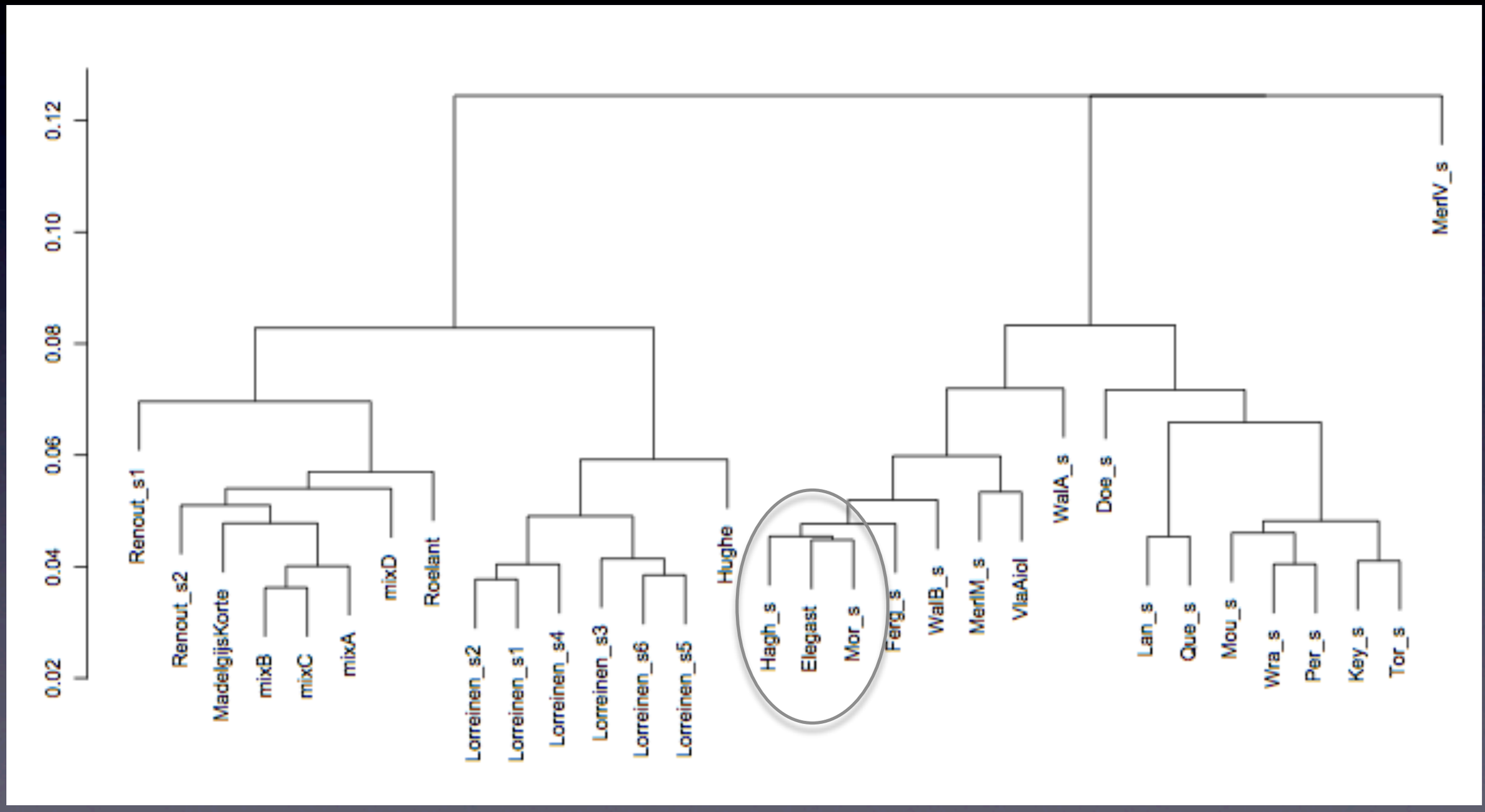
- Koning Artur en Rondetafelridders
- Moriaen, zwarte ridder uit Afrika
- Op zoek naar biologische vader
- Eerste Nederlandse roman met zwart hoofdpersonage
- Ongebruikelijk onderwerp

Icarus…

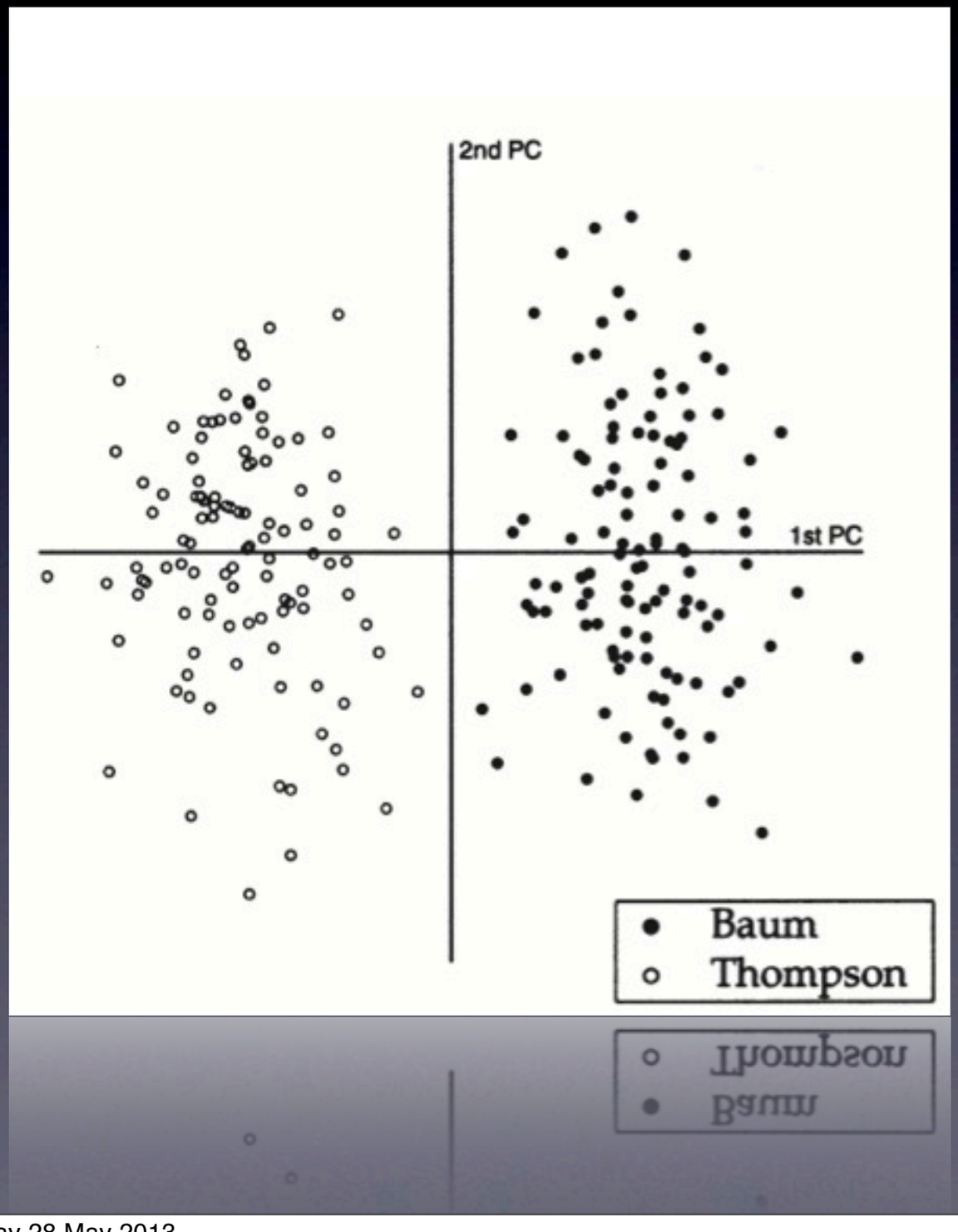


- Toeschrijving afgewezen...
- Té subjectief
- *Ohrenphilologie* ≠ wetenschap
- Bijnaam “Icarus”
- Stylometrie?!

Genres: Serendipiteit?



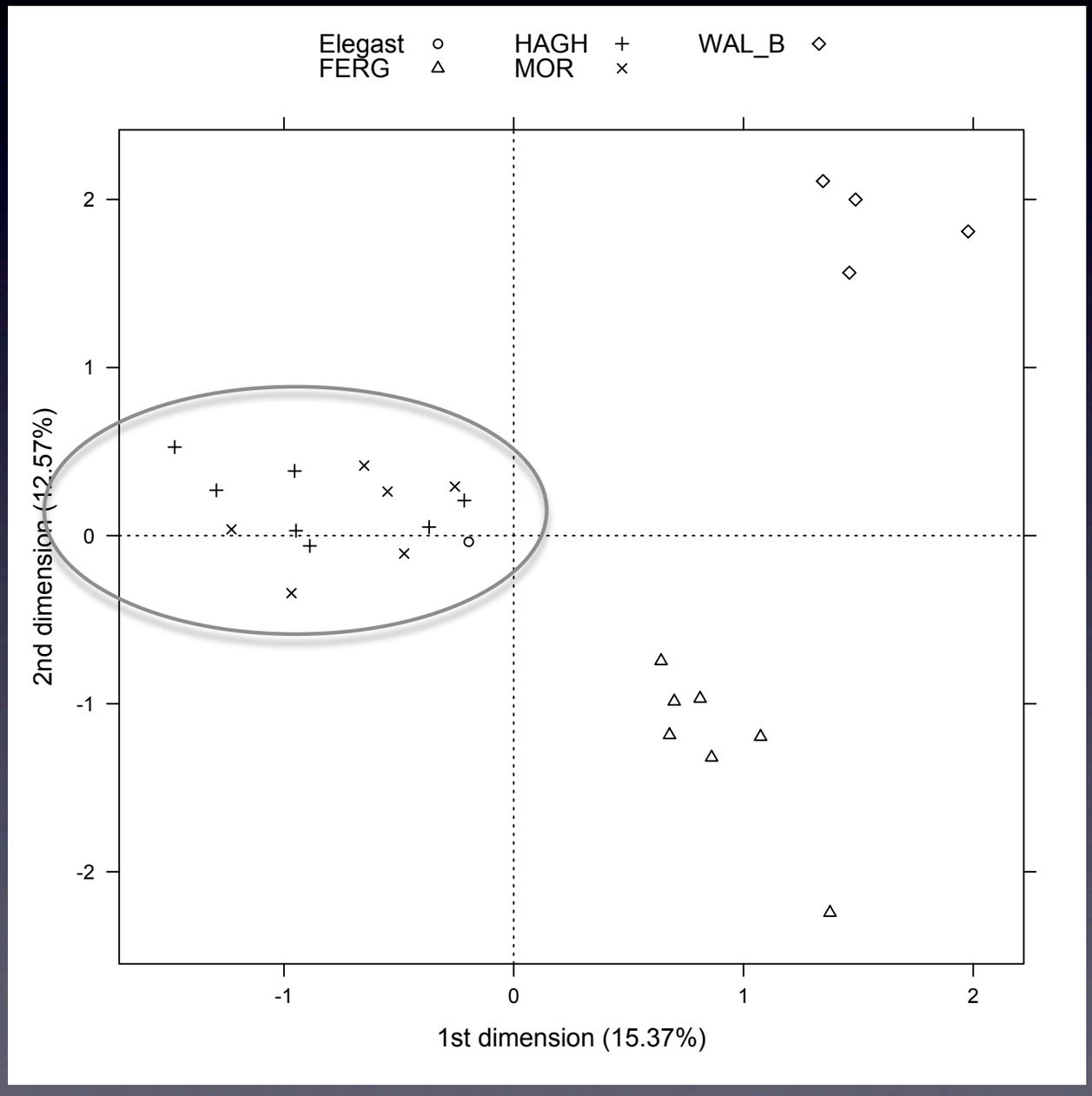
Dimensie-reductie



- Terugdringen # variabelen
- Abstracte, latente dimensies
- Heel populair in stylometrie
- Attributie via clusters
- Bv. PCA, CA, MDS, ...

Eerherstel?

- Correspondentie-analyse
- Sterkst verwant in ridderepiek
- Meting bevestigt buikgevoel
- Eerherstel Heeroma?
- (Maar Gruuthuse...)

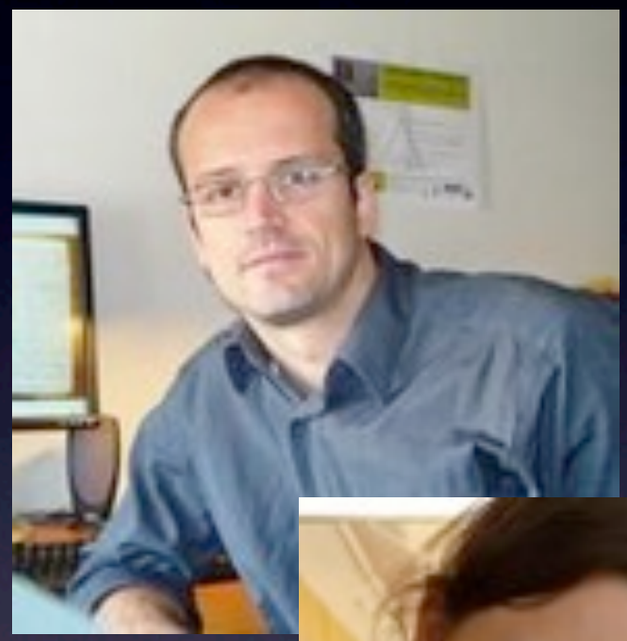


Momenteel...

- FWO-postdoc (2012-2015)
- Uitbreiding stylometrie:
  - proza
  - Latijn
- Veel groter toepassingsgebied

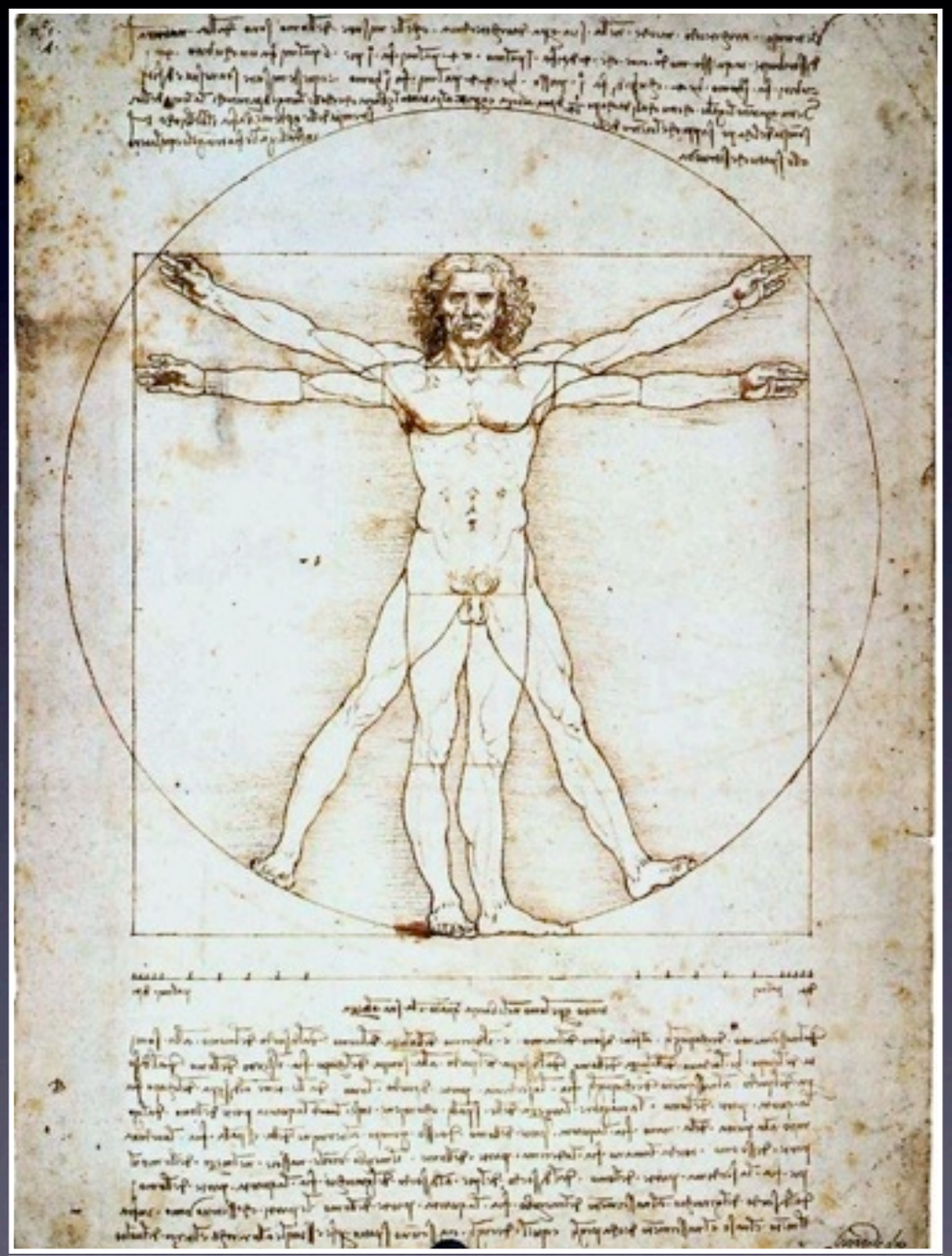
Latijnse mediëvistiek?

- J. Deploige & S. Moens
- Editie 2 korte tekstjes:
  - *Visio de sancto Martino*
  - *Visio ad Guibertum missa*
- Toegeschreven aan Hildegard van Bingen
- Twijfels...



Vitruvische man (*LDO*)

Renaissance 12e eeuw



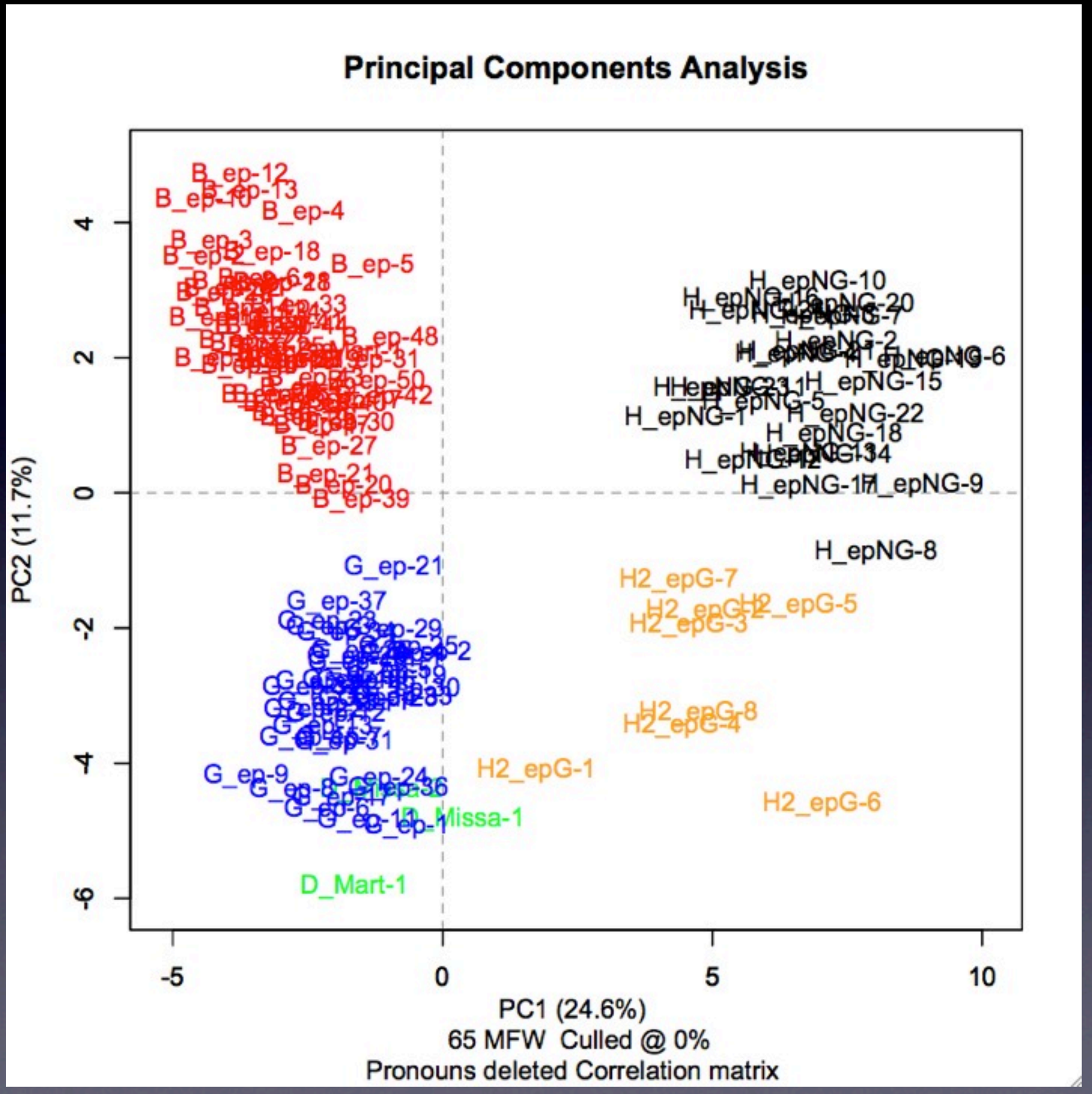
Hildegard von Bingen

- Rijnlandse mystica (1098-1179)
- Hoog aanzien (Bernardus, paus, ...)
- Mystieke visioenen:
  - Dicteerde aan secretarissen
  - Latijn niet volledig machtig (*indocta*)
- Complex "auteurschap"

Laatste secretaris:
Guibert van Gembloux

When you correct [the *Visio de sancto Martino*] you should keep to this rule: that adding, subtracting, and changing nothing, you apply your skill only to make corrections where the order or the rules of correct Latin are violated. Or if you prefer – and this is something I have conceded in this letter beyond my normal practice – you need not hesitate to clothe the whole sequence of the vision in a more becoming garment of speech.

- Stylometrische vingerafdruk
- Toeschrijving onder druk...
- Synergy hypothesis



Meerwaarde stylometrie?

- Erg gerichte applicatie in DH
- Meerwaarde:
  - Falsifiëren oude inzichten
  - Genereren nieuwe inzichten
- Serendipiteit als nevenproduct van *distant reading*

Het gewicht van de auteur

Stylometrische auteursherkenning in Middelnederlandse literatuur

Mike Kestemont

Gent, Koninklijke Academie voor Nederlandse Taal- en Letterkunde, 2013

Toekomst?

- Breder dan stylometrie en mediëvistiek
- *TIME Magazine* archief (i.s.m. Folgert Karsdorp; 1920s-2000s)
- Big Data & Humanities: onderschat probleem

- Recente reeks papers:
  - toptijdschriften
  - “geesteswetenschappelijk”
  - retoriek van de Big Data
  - methodologie unaniem verworpen door vakgenoten

Google Books paper

- *Science* paper
- Google books corpus
- Michel et al.
- “Culturomics”
  - Woordfrequenties
  - Diachroon

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[3] Martin A. Nowak,[2,13,14] Erez Lieberman Aiden[1,2,6,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the underlying phenomena. Attempts to introduce quantitative methods into the study of culture (*1*–*6*) have been hampered by the lack of suitable data.

We report the creation of a corpus of 5,195,769 digitized books containing ~4% of all books ever published. Computational analysis of this corpus enables us to observe cultural trends and subject them to quantitative investigation. 'Culturomics' extends the boundaries of scientific inquiry to a wide array of new phenomena.

The corpus has emerged from Google's effort to digitize books. Most books were drawn from over 40 university libraries around the world. Each page was scanned with custom equipment (*7*), and the text was digitized by means of optical character recognition (OCR). Additional volumes, both physical and digital, were contributed by publishers. Metadata describing the date and place of publication were provided by the libraries and publishers and supplemented with bibliographic databases. Over 15 million books have been digitized [~12% of all books ever published (*7*)]. We selected a subset of over 5 million books for analysis on the basis of the quality of their OCR and metadata (Fig. 1A and fig. S1) (*7*). Periodicals were excluded.

The resulting corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 98 million words per year; by 1900, 1.8 billion; and by 2000, 11 billion (fig. S2).

The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years. The sequence of letters is 1000 times longer than the human genome: If you wrote it out in a straight line, it would reach to the Moon and back 10 times over (*8*).

To make release of the data possible in light of copyright constraints, we restricted this initial study to the question of how often a given 1-gram or *n*-gram was used over time. A 1-gram is a string of characters uninterrupted by a space; this includes words ("banana", "SCUBA") but also numbers ("3.141597") and typos ("excesss"). An *n*-gram is a sequence of 1-grams, such as the phrases "stock market" (a 2-gram) and "the United States of America" (a 5-gram). We restricted *n* to 5 and limited our study to *n*-grams occurring at least 40 times in the corpus.

Usage frequency is computed by dividing the number of instances of the *n*-gram in a given year by the total number of words in the corpus in that year. For instance, in 1861, the 1-gram "slavery" appeared in the corpus 21,460 times, on 11,687 pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is 5.5×10^{-5}. The use of "slavery" peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B).

In contrast, we compare the frequency of "the Great War" to the frequencies of "World War I" and "World War II". References to "the Great War" peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as "World War I" (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cultural phenomena, such as how we remember people and events.

The full data set, which comprises over two billion culturomic trajectories, is available for download or exploration at www.culturomics.org and ngrams.googlelabs.com.

The size of the English lexicon. How many words are in the English language (*9*)?

We call a 1-gram "common" if its frequency is greater than one per billion. [This corresponds to the frequency of the words listed in leading dictionaries (*7*) (fig. S3).] We compiled a list of all common 1-grams in 1900, 1950, and 2000, based on the frequency of each 1-gram in the preceding decade. These lists contained 1,117,997 common 1-grams in 1900, 1,102,920 in 1950, and 1,489,337 in 2000.

Not all common 1-grams are English words. Many fell into three nonword categories: (i) 1-grams with nonalphabetic characters ("3R", "3.141597"), (ii) misspellings ("because", "abberation"), and (iii) foreign words ("sensitivo").

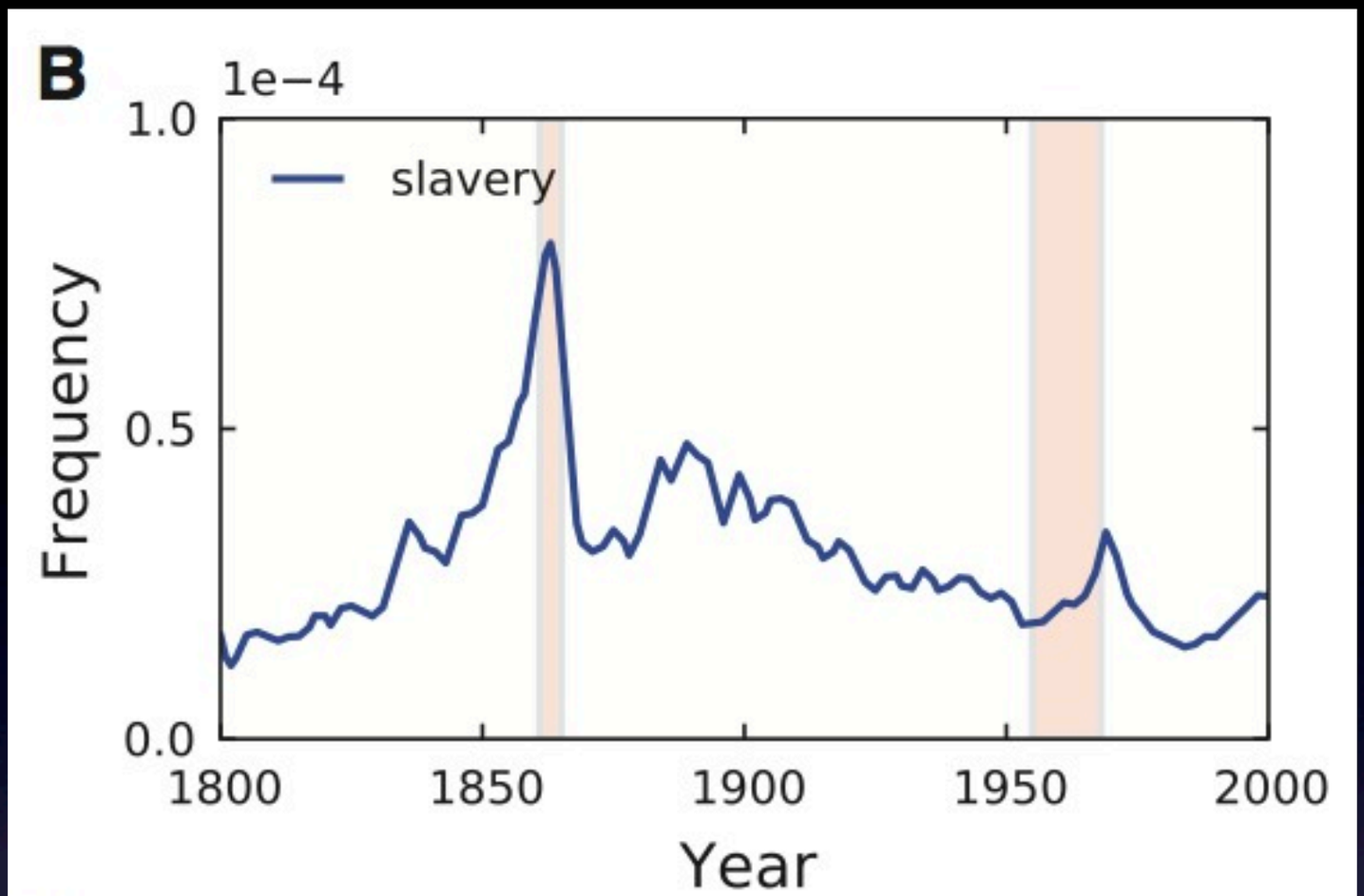
To estimate the number of English words, we manually annotated random samples from the lists of common 1-grams (*7*) and determined what fraction were members of the above nonword categories. The result ranged from 51% of all common 1-grams in 1900 to 31% in 2000.

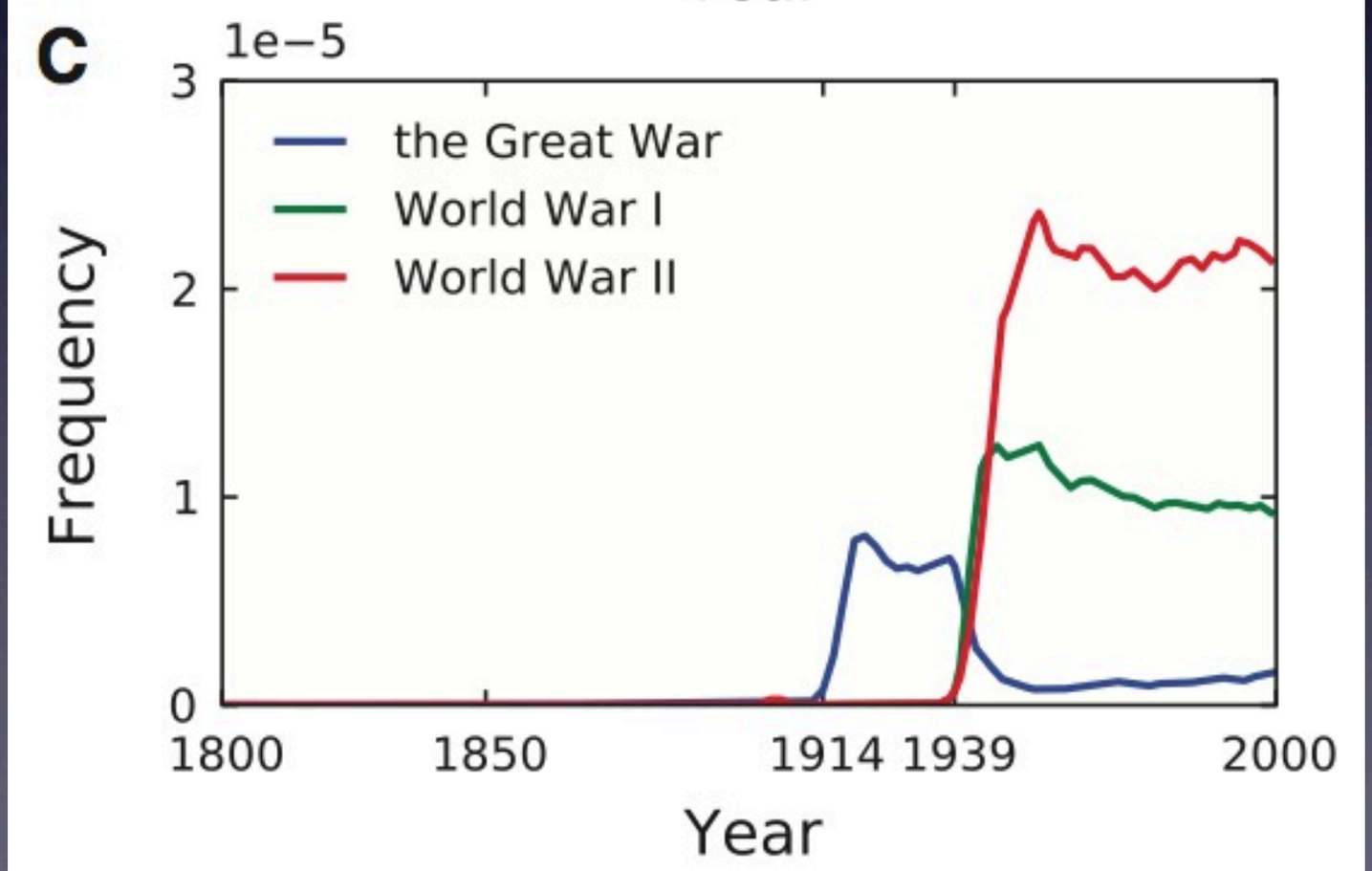
Using this technique, we estimated the number of words in the English lexicon as 544,000 in 1900, 597,000 in 1950, and 1,022,000 in 2000. The lexicon is enjoying a period of enormous growth: The addition of ~8500 words/year has increased the size of the language by over 70% during the past 50 years (Fig. 2A).

Notably, we found more words than appear in any dictionary. For instance, the 2002 *Webster's Third New International Dictionary* (W3), which keeps track of the contemporary American lexicon, lists approximately 348,000 single-word wordforms (*10*); the *American Heritage Dictionary of the English Language, Fourth Edition* (AHD4) lists 116,161 (*11*). (Both contain additional multiword entries.) Part of this gap is because dictionaries often

[1]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA. [2]Cultural Observatory, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA 02138, USA. [4]Department of Psychology, Harvard University, Cambridge, MA 02138, USA. [5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [6]Lieberman-Aiden, Harvard University, Cambridge, MA 02138, USA. [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [8]Harvard Medical School, Boston, MA, 02115, USA. [9]Harvard College, Cambridge, MA 02138, USA. [10]Google, Mountain View, CA 94043, USA. [11]Houghton Mifflin Harcourt, Boston, MA 02116, USA. [12]Encyclopaedia Britannica, Chicago, IL 60654, USA. [13]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [14]Department of Mathematics, Harvard University, Cambridge, MA 02138, USA. [15]Broad Institute of Harvard and MIT, Harvard University, Cambridge, MA 02138, USA. [16]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [17]Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jb.michel@gmail.com (J.-B.M.); erez@erez.com (E.L.A.)





Tikje simplistisch (?),
maar hoe beter doen?

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel[a,b,1], Quentin D. Atkinson[c], Andreea S. Calude[d], and Andrew Meade[a]

[a]School of Biological Sciences, University of Reading, Reading, Berkshire RG6 6AS, United Kingdom; [b]Santa Fe Institute, Santa Fe, NM 87501; [c]School of Psychology, University of Auckland, Auckland 1142, New Zealand; and [d]Linguistics Programme, University of Waikato, Hamilton 3240, New Zealand

Edited* by Colin Renfrew, University of Cambridge, Cambridge, United Kingdom, and approved April 15, 2013 (received for review October 31, 2012)

The search for ever deeper relationships among the World's languages is bedevilled by the fact that most words evolve too rapidly to preserve evidence of their ancestry beyond 5,000 to 9,000 y. On the other hand, quantitative modelling indicates that some "ultraconserved" words exist that might be used to find evidence for deep linguistic relationships beyond that time barrier. Here we use a statistical model, which takes into account the frequency with which words are used in common everyday speech, to predict the existence of a set of such highly conserved words among seven language families of Eurasia postulated to form a linguistic superfamily that evolved from a common ancestor around 15,000 y ago. We derive a dated phylogenetic tree of this proposed superfamily with a time-depth of ∼14,450 y, implying that some frequently used words have been retained in related forms since the end of the last ice age. Words used more than once per 1,000 in everyday speech were 7- to 10-times more likely to show deep ancestry on this tree. Our results suggest a remarkable fidelity in the transmission of some words and give theoretical justification to the search for features of language that might be preserved across wide spans of time and geography.

cultural evolution | phylogeny | historical linguistics

The English word *brother* and the French *frère* are related to the Sanskrit *bhrátr* and the Latin *frāter*, suggesting that words as mere sounds can remain associated with the same meaning for millennia. But how far back in time can traces of a word's genealogical history persist, and can we predict which words are likely to show deep ancestry?

These questions are central to ardent [illegible] tion and to efforts to identify linguistic [illegible] world's languages (1–5). Evidence for [illegible] such as Amerind (6), linking most of the [illegible] New World, and Nostratic (7–9) and Eur[illegible] the major language families of Eurasia. [illegible] identification of putative "cognate" words [illegible] in biology), the sound and meaning corres[illegible] thought to indicate that they derive from a [illegible]

Such evidence is often criticized for [illegible] words are (thought to suffer from too much [illegible] erosion to allow secure identification of [illegible] 5,000 to 9,000 y (11, 12), and second, even [illegible] cognates can be identified, proponents [illegible] ships have been unable to produce statisti[illegible] resemblances they have found are beyond [illegible] by chance between unrelated languages (1[illegible] tests have been used (9, 13), the results [illegible] because of the difficulty of establishing suc[illegible] (hence the number of resemblances expect[illegible]

Both objections can be overcome if it [illegible] a class of words exists whose members' [illegible] correspondences are expected to last long en[illegible] their ancestry between language families [illegible] of years; and (*ii*) these ultraconserved [illegible] a priori and independently of their sou[illegible] other words. Regarding the former, we [illegible] words have about a 50% chance of being replaced by a new monolexic word (a word's linguistic half-life (14, 15)), roughly every 2,000–4,000 y, consistent with the belief that words lose traces of their ancestry quickly. However, some words, such as the numerals, pronouns, and special adverbs (e.g., I, you, here, how, not, there, what, two, five) are replaced far more slowly, with half-lives of once every 10,000, 20,000 or even more years (14, 15).

Usefully, these words can be predicted from information independent of their sounds. We showed in a sample of Indo-European languages that the frequency with which a word is used in everyday speech, along with its part of speech, can predict how rapidly words evolve, with frequently used words on average retained for longer periods of time (14). We have recently extended this result to include speakers from the Uralic, Sino-Tibetan, Niger-Congo, Altaic, and Austronesian families, in addition to Indo-European, plus the isolate Basque and the Creole Tok Pisin (16). Even in languages as widely divergent as these, we found that a measure of the average frequency of use predicted rates of lexical replacement as estimated in the Indo-European languages.

Taken together, these findings suggest that the way we use a core set of vocabulary words in everyday speech is a stable and shared feature of human discourse, and raises the possibility that words will evolve in other language families at rates similar to those found in the Indo-European languages, with frequency of word-use acting as the common causal factor. This provides a statistical framework for predicting—without recourse to sound correspondences—words likely to show deep ancestry among

Ancient symbols, computational linguistics, and the reviewing practices of the general science journals

Richard Sproat*
Center for Spoken Language Understanding

1. Introduction

Few archaeological finds are as evocative as artifacts inscribed with symbols. Whenev[illegible] an archaeologist finds a potsherd or a seal impression that seems to have symbo[illegible] scratched or impressed on the surface, it is natural to want to "read" the symbols. A[illegible] if the symbols come from an undeciphered or previously unknown symbol system[illegible] is common to ask what language the symbols supposedly represent and whether t[illegible] system can be deciphered.

Of course the first question that really should be asked is whether the symbols a[illegible] writing system, as linguists usually define it, is a symbol system th[illegible] language. Familiar examples are alphabets such as the Latin, Gree[illegible] alphabets, alphasyllabaries such as Devanagari or Tamil, syllabari[illegible] e Kana, and morphosyllabic systems like Chinese characters. B[illegible] t do not encode language abound: European heraldry, mathematic[illegible] ion (used to represent dance), and boy scout merit badges are [illegible] l systems that represent things, but do not function as part of [illegible] nts language.

known system is writing or not is a difficult question to answe[illegible] ered definitively in the affirmative if one can develop a verifiab[illegible] some language or languages. Statistical techniques have been used [illegible] ears, but these have always been used under the assumption that t[illegible] g with *is* writing, and the techniques are used to uncover patterns [illegible] ght aid in the decipherment. Patterns of symbol distribution mig[illegible] bol system is *not* linguistic: for example, odd repetition patter[illegible] n that a symbol system is unlikely to be writing. But until recent[illegible] l that statistical techniques could be used to determine that a syste[illegible]

e quite a surprise when, in April 2009, there appeared in *Scie*[illegible] tajesh Rao of the University of Washington and colleagues at tw[illegible]

nguage Understanding, Oregon Health & Science University, 20000 NW Walker R[illegible], USA. E-mail: rws@xoba.com.

existence of quasi-Zipfian distributions in symbol systems to argue for their statu[illegible] s figure in the work of Rao and colleagues. But since it has been long known that [illegible] hold of many things besides language, such arguments are easy to dismiss.

Computational Linguistics

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5,6]† Yuan Kui Shen,[2,4,7] Aviva Presser Aiden,[2,4,8] Adrian Veres,[2,4,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[3] Martin A. Nowak,[2,13,14] Erez Lieberman Aiden[1,2,4,14,15,16,17]†

[1]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA. [2]Cultural Observatory, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA 02138, USA. [4]Department of Psychology, Harvard University, Cambridge, MA 02138, USA. [5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [6]Edmond and Lily Safra Center for Brain Sciences, Harvard University, Cambridge, MA 02138, USA. [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [8]Harvard Medical School, Boston, MA, 02115, USA. [9]Harvard College, Cambridge, MA 02138, USA. [10]Google, Mountain View, CA 94043, USA. [11]Houghton Mifflin Harcourt, Boston, MA 02116, USA. [12]Encyclopædia Britannica, Chicago, IL 60654, USA. [13]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [14]Department of Mathematics, Harvard University, Cambridge, MA 02138, USA. [15]Broad Institute of Harvard and MIT, Harvard University, Cambridge, MA 02136, USA. [16]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [17]Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.
(To whom correspondence should be addressed. E-mail: jbmichel@gmail.com (J.-B.M.); erez@erez.com (E.L.A.)

Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the underlying phenomena. Attempts to introduce quantitative methods into the study of culture (*1*–*6*) have been hampered by the lack of suitable data.

We report the creation of a corpus of 5,195,769 digitized books containing ~4% of all books ever published. Computational analysis of this corpus enables us to observe cultural trends and subject them to quantitative investigation. 'Culturomics' extends the boundaries of scientific inquiry to a wide array of new phenomena.

The corpus has emerged from Google's effort to digitize books. Most books were drawn from over 40 university libraries around the world. Each page was scanned with custom equipment (*7*), and the text was digitized by means of optical character recognition (OCR). Additional volumes, both physical and digital, were contributed by publishers. Metadata describing the date and place of publication were provided by the libraries and publishers and supplemented with bibliographic databases. Over 15 million books have been digitized (~12% of all books ever published (*7*)). We selected a subset of over 5 million books for analysis on the basis of the quality of their OCR and metadata (Fig. 1A and fig. S1) (*7*). Periodicals were excluded.

The resulting corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 98 million words per year; by 1900, 1.8 billion; and by 2000, 11 billion (fig. S2).

The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years. The sequence of letters is 1000 times longer than the human genome: If you were it out in a straight line, it would reach to the Moon and back 10 times over (*8*).

To make release of the data possible in light of copyright constraints, we restricted this initial study to the question of how often a given 1-gram or *n*-gram was used over time. A 1-gram is a string of characters uninterrupted by a space; this includes words ("banana", "SCUBA") but also numbers ("3.14159") and typos ("excesss"). An *n*-gram is a sequence of 1-grams, such as the phrases "stock market" (a 2-gram) and "the United States of America" (a 5-gram). We restricted *n* to 5 and limited our study to *n*-grams occurring at least 40 times in the corpus.

Usage frequency is computed by dividing the number of instances of the *n*-gram in a given year by the total number of words in the corpus in that year. For instance, in 1861, the 1-gram "slavery" appeared in the corpus 21,460 times, on 11,687 pages of 1208 books. The corpus contains 386,434,758 words from 1861, thus, the frequency is 5.5×10^{-5}. The use of "slavery" peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B).

In contrast, we compare the frequency of "the Great War" to the frequencies of "World War I" and "World War II". References to "the Great War" peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as "World War I" (Fig. 1C).

These examples highlight two central factors that contribute to cultural trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cultural phenomena, such as how we remember people and events.

The full data set, which comprises over two billion culturomic trajectories, is available for download or exploration at www.culturomics.org and ngrams.googlelabs.com.

The size of the English lexicon. How many words are in the English language (*9*)?

We call a 1-gram "common" if its frequency is greater than one per billion. [This corresponds to the frequency of the words listed in leading dictionaries (*7*) (fig. S3).] We compiled a list of all common 1-grams in 1900, 1950, and 2000, based on the frequency of each 1-gram in the preceding decade. These lists contained 1,117,997 common 1-grams in 1900, 1,102,820 in 1950, and 1,489,337 in 2000.

Not all common 1-grams are English words. Many fall into three nonword categories: (*i*) 1-grams with nonalphabetic characters ("3b", "3.14159"), (*ii*) misspellings ("because", "abberation"), and (*iii*) foreign words ("sensitivo").

To estimate the number of English words, we manually annotated random samples from the lists of common 1-grams (*7*) and determined what fraction were members of the above nonword categories. The result: ranged from 51% of all common 1-grams in 1900 to 31% in 2000.

Using this technique, we estimated the number of words in the English lexicon as 544,000 in 1900, 597,000 in 1950, and 1,022,000 in 2000. The lexicon is enjoying a period of enormous growth: The addition of ~8500 words/year has increased the size of the language by over 70% during the past 50 years (Fig. 2A).

Notably, we found more words than appear in any dictionary. For instance, the 2002 *Webster's Third New International Dictionary* (W3), which keeps track of the contemporary American lexicon, lists approximately 348,000 single-word-wordforms (*10*), the *American Heritage Dictionary of the English Language, Fourth Edition* (AHD4) lists 116,161 (*11*). (Both contain additional multiword entries.) Part of this gap is because dictionaries often

Quantitative patterns of stylistic influence in the evolution of literature

James M. Hughes[a], Nicholas J. Foti[b], David C. Krakauer[a,c], and Daniel N. Rockmore[a,b,d,1]

[a]Department of Computer Science, Dartmouth College, Hanover, NH 03755; [b]Santa Fe Institute, Santa Fe, NM 87501; [c]Wisconsin Institute for Discovery, University of Wisconsin, Madison, WI 53715; [d]Department of Mathematics, Dartmouth College, Hanover, NH 03755; and [e]Neukom Institute for Computational Science, Dartmouth College, Hanover, NH 03755

Edited by* Michael S. Gazzaniga, University of California, Santa Barbara, Santa Barbara, CA, and approved March 13, 2012 (received for review September 21, 2011)

Literature is a form of expression whose temporal structure, both in content and style, provides a historical record of the evolution of culture. In this work we take on a quantitative analysis of literary style and conduct the first large-scale temporal stylometric study of literature by using the vast holdings in the Project Gutenberg Digital Library corpus. We find temporal stylistic localization among authors through the analysis of the similarity structure in feature vectors derived from content-free word usage, nonhomogeneous decay rates of stylistic influence, and an accelerating rate of decay of influence among modern authors. Within a given time period we also find evidence for stylistic coherence with a given literary topic, such that writers in different fields adopt different literary styles. This study gives quantitative support to the notion of a literary "style of a time" with a strong trend toward increasingly contemporaneous stylistic influence.

cultural evolution | stylometry | culture | complexity | big data

Written works, or literature, provide one of the great bodies of cultural artifacts. The analysis of literature typically involves the aggregation of information on several levels, ranging from words to sentences and even larger scale properties of temporal narratives such as structure, plot, and the use of irony and metaphor (*1*–*3*). Quantitative methods have long been applied to literature, most notably in the analysis of style, which can be traced back to a comment by the mathematician Augustus de Morgan regarding the attribution of the Pauline epistles (*4*) and the late nineteenth century work of the historian of philosophy Wincenty Lutasłowski, who brought basic statistical ideas of word usage to the problem of dating the dialogues of Plato (*5*). It was Lutasłowski who coined the word "stylometry" to describe such an approach to investigating questions of literary style. Since then, a wide range of statistical techniques have been developed toward this end (*6*), generally with the goal of settling questions of author attribution (see, e.g., refs. *6*–*11*). Stylometric studies have also been pursued in the study of visual art (*12*, *13*) and music [both in composition (*14*–*16*) and performance (*17*)], and are part of a growing body of work in the quantitative analysis of cultural artifacts (*18*).

In this paper we report our findings from the first large-scale stylometric analysis of literature. The goal of this work is not author attribution—for the authorship of all the works is well known—but is instead to articulate, in a quantitative fashion, large-scale temporal trends in literary (i.e., writing) style. This type of study has been, until now, impossible to undertake, but the advent of mass digitization has created dramatic new opportunities for scholarly studies in literature as well as in other disciplines (*19*). Our literature sample is obtained from the Project Gutenberg Digital Library (http://www.gutenberg.org/wiki/Gutenberg:About). Project Gutenberg consists of more than 30,000 public domain texts, music, audiobooks, etc., is freely available online, and is among the digital archives that have become, over the past 60 yr, crucial components of the preservation of cultural artifacts (*18*).

In scope, our work is related to, but quite different from, recent studies in the dating of literary works (*20*), the analysis of the coarse-grained structure of literary history (and the evolution of genre) (*21*), and most notably, a recent analysis of Google Books (*22*), wherein the temporal trends in content-word usage were articulated. Content words also form the basis of the various topic model analyses that have been applied to large corpuses of science texts (see, e.g., ref. *23*).

In contrast, the work presented here focuses on the usage of content-free words as the basis of the first large-scale study of the similarity structure of literary style. Content-free words are the "syntactic glue" of a language: They are words that carry little meaning on their own but form the bridge between words that convey meaning. Their joint frequency of usage is known to provide a useful stylistic fingerprint for authorship (*8*, *11*), and thus suggests a method of comparing author styles. When we consider content-free word frequencies from a large number of authors and works over a long period of time, we can ask questions related to temporal trends in similarity. The primary results of our analysis are that time provides the most coherent means of clustering work and a trend of diminishing stylistic influence as we move forward in time. Such a finding is consistent with a simple evolutionary model for stylistic influence, which assumes that imitation attends preferentially to contemporary authors. In addition, we uncover quantitative support of the previously purely anecdotal notion of a literary "style of a time." Taken together, these two findings suggest the utility and perhaps the creation of a new field of stylometric analysis in culturomics.

Materials and Methods

In our experiments, we studied a subset of the authors in the Project Gutenberg database composed of those who wrote after the year 1500, had at least five works in English in the Project Gutenberg collection, and for whom we had birth and death date information. This left us with 537 authors. For each author, we created a representative feature vector by aggregating the content-free word frequencies for each individual work by that author. In total, we analyzed 7,710 works.

In our experiments, we used a list of 307 content-free words that included prepositions, articles, conjunctions, "to be" verbs, and some common nouns and pronouns (see Table S1 for a complete list). We did not attempt to semantically disambiguate between occurrences of homographs in situ (e.g., when using "to" as a preposition or as an indicator of an infinitive verb). Doing so would require a sophisticated grammatical model, and it was not our aim to model this particular aspect of word usage. We believe that ignoring these distinctions is not likely to greatly affect our results, because words that account for the greatest differences in usage frequency among

Author contributions: J.M.H., N.J.F., D.C.K., and D.N.R. designed research; J.M.H., N.J.F., D.C.K., and D.N.R. performed research; J.M.H., N.J.F., D.C.K., and D.N.R. contributed new reagents/analytic tools; J.M.H., N.J.F., D.C.K., and D.N.R. analyzed data; and J.M.H., N.J.F., D.C.K., and D.N.R. wrote the paper.

The authors declare no conflict of interest.
*This Direct Submission article had a prearranged editor.

To whom correspondence should be addressed. E-mail: rockmore@dartmouth.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1115407109/-/DCSupplemental.

"Would a paper that made some blatantly wrong claim about genetics be published in such venues?" (R. Sproat)

Discussie

- Goed voor visibiliteit vakgebied
- Kruisbestuiving exacte wetenschappen
- Afgunst speelt mee?

- Maar ook beschermen eigen vakgebied:
  - Retoriek Big Data wordt ‘misbruikt’...
  - Methodologie unaniem afgewezen...
  - Onderzoekers zijn allesbehalve geesteswetenschappers...
- Hebben de Geesteswetenschappen een eigen ‘toptijdschrift’ nodig?