The Weight of the Author
Quantitative Authorship Attribution in Medieval Dutch Literature

Mike Kestemont (UA/FWO) | 9 May 2012 | Nijmegen | MPI (LTA 1)
Supervisors: Frank Willaert (UA | ISLN) & Walter Daelemans (UA | CLiPS)
The Death of the Author?

- R. Barthes (1968)
- La mort de l’auteur
- “It doesn’t matter who wrote a text”
- Revolutionary idea
- Derrida, Kristeva, Foucault, ...
- Influential discourse
Yet this erasure of the author has remained paradoxical, not least because it was powered by the creativity of a number of profoundly individualistic thinkers and writers. There was never any doubt as to where royalty cheques for Barthes, Foucault, Lacan, Kristeva and Derrida ought to be sent: they reasserted heroic authorship even in their questioning it (Love 2002: 7).

Ironically...
“Author-centric” society
• Fascination for author
• Lack of info
• Lots of **anonymous** texts
  • *humilitas* (modesty?)
  • material damage
  • small-scale literature
• Curiosity & Frustration

**Medieval studies**
Middle Dutch Literature (ca. 1200-1400)
Style-Based Authorship Attribution

- Big issue historically
- A lot of speculation
- Manual, subjective research
- “To me, it seems like …”, “It could be that”, “Plausibly, ...”
- Problem: only text...

---
“A ready-made statistical method for the stylistic comparison of two texts does not (yet) exist, but I wonder whether it will ever be possible to develop such a method. It is clear that literary texts are no quantitative units that can be compared.”

• Representative quote (1999)
• A lot of digitized texts (Cd-rom Middelnederlands 1996):
  • Distrust about (later) Digital Humanities
  • Misinformed... Methods did exist at the time!
• **Stylometry** – quantitative study of writing style (Holmes 1998)

• “by **measuring some textual features**, we can distinguish between texts written by different authors.” (Stamatatos 2009: 538)

• Basic “**fingerprint**” assumption: each author has unique writing style

• Strong formulation: **Stylome Hypothesis** (Van Halteren et al. 2005); cf. genome

**Authorship Attribution**
Stylometric Authorship studies

- Young paradigm (1960s)
- Mosteller & Wallace (USA)
- Federalist papers (1780s)
- By “Publius”...

Methodological novelties:

1. Quantitative measurements &gt;&lt; traditional “taste-based methodologies”
2. New features
## New feature type

<table>
<thead>
<tr>
<th>Tradition</th>
<th>Mosteller &amp; Wallace</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Manual analysis (<em>close reading</em>)</td>
<td>• Quantitative analysis (<em>stats</em>)</td>
</tr>
<tr>
<td>• Focus on features <strong>catching</strong> the reader’s eye</td>
<td>• Focus on inconspicuous features</td>
</tr>
<tr>
<td>• <strong>Conspicuous</strong> nouns, rare verbs, ...</td>
<td>• <strong>Function words or functors</strong> – articles, prepositions, pronouns, ...</td>
</tr>
<tr>
<td>• Checklist rare items with low frequency</td>
<td>• Extremely <strong>high frequency</strong> items</td>
</tr>
<tr>
<td>• <strong>But</strong> school formation, forgery, imitation, content specific, ...</td>
<td>• Why?</td>
</tr>
</tbody>
</table>
• High frequency: reliable stats
  • All authors use them
• Relatively content-independent

Function words  (Binongo 2003)
Count the number of $f$'s on the following slide...
Finished files are the result of years of scientific study combined with the experience of many years.
Finished files are the result of years of scientific study combined with the experience of many years.

‘Of unconsciously processed?’ (Schindler 1978)
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 typescases or a list, rather than a paragraph, format. These effects for the word and were not found for the control word set. These results were taken as evidence that familiar...
• “Functors” popular in stylometry
• Relative frequencies of functors
• Input for algorithms
• Statistics and Machine learning
• The Secret Life of Pronouns
  (Pennebaker 2011 – trauma writing)

Function words
Parallel in art history

G. Morelli (1816–1891)
• Goal of PhD
• Apply to Middle Dutch?
• Ideal algorithm: “Text-in-Author-Out”
• No real computational tradition yet in MDS...
• Fascinating: apply rocket science technologies to ancient texts!
• Today no summary, but “story”

Medieval texts?
Scribes

- Medieval texts?
- Manu-scripts (parchment)
- No printing press yet
- Each copy: “hard copy”
- Handwritten by scribes
- Each copy was unique
• No standard language or spelling yet...
• Each scribe adapted text to own:
  • spelling conventions
  • local dialect
  • stylistic preferences
• Copies often deviate from exemplar
• Original author’s text?

Language? (Van Dalen-Oskam & Van Zundert 2007)
Huge variation between MSS of “same text”
<table>
<thead>
<tr>
<th>Handschrift A</th>
<th>Handschrift D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample $A_1$</td>
<td>Sample $D_1$</td>
</tr>
<tr>
<td>Nota Tier stont ende ter seluer uren</td>
<td>Ter stont ende ter seluer vren</td>
</tr>
<tr>
<td>So was van allen creaturen</td>
<td>was van allen creaturen</td>
</tr>
<tr>
<td>Dat uerstaet van allen beesten</td>
<td>Dat verstaet van allen beesten</td>
</tr>
<tr>
<td>Tserpent lustechst na der Jeesten</td>
<td>serpent lustech na der ieesten</td>
</tr>
<tr>
<td>Ende harde scal oec van sinne</td>
<td>Ende herde scal oec van sinne</td>
</tr>
<tr>
<td>Sample $A_2$</td>
<td>Sample $D_2$</td>
</tr>
<tr>
<td>Want die duuel waser Inne</td>
<td>want die duuel waser inne</td>
</tr>
<tr>
<td>bedi dat lucifer die scone</td>
<td>bedi dat lucifer die scone</td>
</tr>
<tr>
<td>Was geworpen uten trone</td>
<td>was geworpen uten trone</td>
</tr>
<tr>
<td>hadde hijs nijt in alre wijs</td>
<td>had hijs nijt in alre wijs</td>
</tr>
<tr>
<td>Dat adam was int paradijs</td>
<td>dat adam was int paradijs</td>
</tr>
<tr>
<td>Sample $A_3$</td>
<td>Sample $D_3$</td>
</tr>
<tr>
<td>Ende pensde hoe dat hine uut dade steken</td>
<td>Ende pensde hoe dat hine vt dade steken</td>
</tr>
<tr>
<td>Dade hi hem tgebod gods breken</td>
<td>dadi hem tgebod gods breken</td>
</tr>
<tr>
<td>So moeste hire ute pensde hi dan</td>
<td>so moeste hire vte pensdi dan</td>
</tr>
</tbody>
</table>

Example Rijmbijbel, 13th c. (Kestemont & Van Dalen-Oskam 2009)
Principal components analysis

- **PCA**: established technique in stylometry
- Applied to frequencies of function words in texts
- Visualizes the stylistic variation in texts
- Often yields clusters of samples by same author (Binongo 2003: 13)

Figure 5. Baum vs. Thompson.
Parallel manuscripts of “same text”: similar author separation effect!
Huge “stylistic” variation between manuscripts of “same text”

- Considerable problems for stylometry
- Texts by same author but different scribe
  - Artificially *distant* in style?
- Texts by different author but same scribe:
  - Artificially *close* in style?

<table>
<thead>
<tr>
<th>D</th>
<th>Ter stont ende ter seluer vren</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Tier stont ende ter seluer vren</td>
</tr>
<tr>
<td>F</td>
<td>Tlere stont enter seluer vren</td>
</tr>
<tr>
<td>G</td>
<td>Tottien stonden en ter uren</td>
</tr>
<tr>
<td>H</td>
<td>TEn stonden ende ter seluer vren</td>
</tr>
<tr>
<td>I</td>
<td>Tjerst stont ende tier veren</td>
</tr>
<tr>
<td>J</td>
<td>Tyer stont ende tier seluer vren</td>
</tr>
<tr>
<td>N</td>
<td>TJer stont tier seluer vre</td>
</tr>
</tbody>
</table>
• Solution?
• (Nearly) all Middle Dutch poetry rhymed
• Couplets (aabbccdd...)
• Very difficult to change
• Skeleton of text
• **Bypass** for scribes

**Bypass:** *rhyme words*
• Still spelling variation in rhyme position
• Solution: lemmatize rhyme words for normalization
• Use frequencies of lemmas as input for algorithms

Lemmatization (Kestemont et al. 2010)
• Obs: “Always the same rhyme words”
• Finding rhyme words can be difficult
• So poets rely on formulaic rhyme patterns ("stopgaps")
• Cf. pop songs: “The sky is blue : I love ...”
• **High-frequency “stopgaps”** in Middle Dutch
• As surrogate for high-frequency functors in modern texts
Case study *Spiegel historiael*

- Excellent test case?
- “Mirror of History”
- **Gigantic** Middle Dutch Rhymed Chronicle
- World history from Genesis until “present” (ca. 1300)
- 3 consecutive authors: *Maerlant, Utenbroeke en Velthem*
- 5 parts in total
• Maerlant (P3) & Utenbroeke (P2) close to each other
• Dialect, topic workplace, schooling, ...
• Difficult for style-based separation?
• Perfect case for authorship authorship attribution
• Machine Learning
• An “intelligent”, self-learning computer
• E.g. spam filtering
• Learn how to assign labels to texts
• “Classifier”
• Based on word frequencies
<table>
<thead>
<tr>
<th></th>
<th>“Lottery”</th>
<th>“Thesis”</th>
<th>“Viagra”</th>
<th>Label?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example mail 1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>SPAM</td>
</tr>
<tr>
<td>Example mail 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NOT SPAM</td>
</tr>
<tr>
<td>Example mail 3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>NOT SPAM</td>
</tr>
<tr>
<td>Example mail 4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>SPAM</td>
</tr>
</tbody>
</table>

**Classic example (“training data”)**
• Learn “authorship label” using simple classifier
• Training data vs. test data
• Leave-one-out
• Accuracy? Influence of specific parameters?

Apply to authorship
Works for Maerlant & Utenbroeke?
Encouraging results for two authors

But what about a third author?

Lodewijk van Velthem in P5 ("Fifth Part")

= 7 books (‘contemporary history ca. 1300)
Good results, however...

Book 4 was always wrongly attributed...
Battle of the Golden Spurs (1302)
Battle of the Golden Spurs (1302)

- Flemish yearly holiday
- 11 July 1302
- Most celebrated passage from Velthem’s oeuvre
- Extremely detailed account
- But was it written by Velthem?
Internal Plagiarism Detection
“Plagiarism?”

- Traditional investigation for other clues
- Awkward:
  - Suspiciously detailed...
  - Very long
  - Unusual metaphors
  - Other dialect
  - Velthem often “plagiarized”
  -...
- Did he “rip” this text from someone else?
- Written by Flemish eye witness shortly after 1302?
• **Icon** of Middle Dutch Literature
• 13\(^{\text{th}}\) c. Flanders, **anonymous**
• Short, epic story in rhymes
• *Fraeye historie ende al waer*
• Read in most high schools
  • Charlemagne awakened by angel
  • “Go out and steal!”
  • Meets the magician-thief Elegast
  • Discovers conspiracy against life

**Karel ende Elegast**
• Professor of Dutch literature
• Poet himself
• Years before his death
• Middle Dutch literature
• “Subjective” stylistic analysis
• Controversial research
• “Heard” the same voice in *Elegast* and *Moriaen*
• Same author?!

**K.H. Heeroma (1909-1972)**
• Story about King Arthur and Knights of Round Table
• Knights meet Moriaen, a black-skinned knight from Africa
• Looking for his biological father
• First ‘novel’ in Dutch literature with black protagonist
• Extremely unconventional at the time
• Anonymous
• Cf. stealing king...

Roman van Moriaen
• Attribution rejected by researchers
• Too subjective, no evidence
• “Hearing” ≠ scientific
• Heeroma = “Icarus” of Dutch studies
• What about stylometry?

Same author?
Cluster analysis of entire corpus...
• Correspondence-analysis
• Closer than any other Middle Dutch epics...
• “Objective method” confirms “subjective” guess
• Rehabilitation of Heeroma?

Rehabilitation for Icarus?
Mike Kestemont nodigt u uit op de openbare verdediging van zijn proefschrift

Het gewicht van de auteur
Een onderzoek naar stylometrische auteursherkenning in de Middelnederlandse epiëk

Promotoren: Prof. dr. Frank Willaert en Prof. dr. Walter Daelemans

23 mei 2012 om 15u, met aansluitend een receptie
Hof van Liere, F. de Tassiszaal
Prinsstraat 13, 2000 Antwerpen

Gelieve uw komst te bevestigen vóór 16 mei 2012 (Mike.Kestemont@ua.ac.be).

Some references / www.mike-kestemont.org