INVESTIGATING THE /ING-FORM NETWORK
IN THE IDIOLECTS OF 17TH CENTURY AUTHORS

Lauren Fonteyn
Andrea Nini

University of Manchester
ICCG-10 - Sorbonne Paris Cité University (USPC) - July 16th-20th
The scientific study of language as applied to forensic purposes and contexts

(McMenamin 2002)
Forensic Linguistics

- Study of the language of the law
- Study of language as evidence
- Disputed authorship
Relative frequency

- a = 3%
- an = 4%
- all = 0.3%
- also = 0.1%
- should = 0.02%
- some = 0.001%
- such = 1%
- of = 6%
- now = 0.2%
- upon = 0.1%

**AUTHOR A**

Relative frequency

- a = 2%
- an = 1%
- all = 0.2%
- also = 1%
- should = 0.002%
- some = 0.1%
- such = 2%
- of = 0.02%
- now = 0.1%
- upon = 0.01%

**AUTHOR B**
### Requirements
- Most frequent 150 word types
- About 1000-1500 word tokens per text

### Findings
- 85% of the times the real author is among the top 5 candidates

### Explanation
- “How is it that such a primitive statistical instrument can satisfy these purposes? The answer must lie, I believe, in areas where we are still extremely ignorant—in the communicative resilience of the language and the **astonishing force of human individuality**.” (Burrows 2002)

The Delta (Δ) score:

\[ \Delta_{(AB)} = \frac{1}{n} \sum \left| \frac{(A - \mu)}{\sigma} - \frac{(B - \mu)}{\sigma} \right| \]
“The totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker is an idiolect.”

Bloch (1948)
“Every speaker has a very large active vocabulary built up over many years, which will differ from the vocabularies others have similarly built up, not only in terms of actual items but also in preferences for selecting certain items rather than others. Thus, whereas in principle any speaker/writer can use any word at any time, speakers in fact tend to make typical and individuating co-selections of preferred words. This implies that it should be possible to devise a method of linguistic fingerprinting—in other words that the linguistic ‘impressions’ created by a given speaker/writer should be usable, just like a signature, to identify them.”

Coulthard (2004)
<table>
<thead>
<tr>
<th>“STRAWBERRY” Nos. 1-3</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. STRAWBERRY</td>
<td>2. STRAWBERRY</td>
<td>3. STRAWBERRY</td>
<td></td>
</tr>
<tr>
<td>“STRAWBERRIES” Nos. 4-23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. STRAWBERRIES</td>
<td>5. STRAWBERRIES</td>
<td>6. STRAWBERRY’S</td>
<td></td>
</tr>
<tr>
<td>7. STRAWBERRIES</td>
<td>8. STRAWBERRY’S</td>
<td>9. STRAW-BERRIES</td>
<td></td>
</tr>
<tr>
<td>10. BoysenBerries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“CALIFORNIA STRAWBERRIES”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. STRAWBERRIES</td>
<td>12. STRAWBERRY</td>
<td>13. STRAWBERRY IES</td>
<td></td>
</tr>
<tr>
<td>14. STRAWBERRY</td>
<td>15. STRAWBERRY</td>
<td>16. STRAWBERRY</td>
<td></td>
</tr>
<tr>
<td>17. STRAWBERRY</td>
<td>18. STRAWBERRY-S</td>
<td>19. STRAWBERRY-S</td>
<td></td>
</tr>
</tbody>
</table>


- **Spelling**
  - Berry
  - Bery
  - Berries

- **Morphology**
  - Berrie’s
  - Berrys

- **Word formation**
  - Strawberry
  - Strawberries

- **Positioning**
  - Strawberries

---

GERUNDS & IDIOLECT
THE GERUND ‘ALTERNATION’

- Deverbal nominalization in -ing

- Two types:
  1. **nominal gerund**: … the dishonour of Gods Name should affect us more then *the shedding of the warmest blood in our veins* (John Flavell, 1668)
  2. **verbal gerund**: he also made an end of … ø *Shedding ø the Blood of Rams, Lambs, Heifers, Goats and other Creatures* for the Sins and Transgressions of Men (George Fox, 1686)
HISTORICAL DEVELOPMENT

Old English
Gerund is an abstract deverbal noun, with nominal syntactic features (NG) (e.g. by writing of a letter)

Middle English
Gerund was re-analysed as part of the verb system and acquired the ability to govern a direct object (c. 1250 - e.g. by writing a letter)

Modern English
Gradual spread of the verbalized gerund (Fanego 2004)
HISTORICAL DEVELOPMENT

- Competition between two (or more) forms in the language system has either one of two outcomes (Traugott & Trousdale 2013: 18): **substitution** or **retention** (development towards division of labour);

- Historical process described as **substitution of NG by VG** in Middle and (Modern English (De Smet 2008; Nevalainen et al. 2011);

- Gradual reorganisation of the English ‘*ing*-form network’ – the functionally hybrid gerund splits into a nominal and a clausal component (Fonteyn forthc.).

(Figure taken from Fonteyn (2017) – data from PPCEME and PPCMBE)
“Some subtypes of the new construction [i.e. VG] became possible before others, their generalization being largely governed by two different linguistic hierarchies” (Fanego 2004: 50)

HISTORICAL DEVELOPMENT

• hierarchy of relative ‘nominality’
  (1) ø (by) eating the forbidden fruit
  (2) poss (by) Adam’s eating the forbidden fruit
  (3) the (by) the / an unadvised eating the forbidden fruit

• grammatical relations hierarchy
  (1) prep - by my not doing it
  (2) object - It does not excuse my not doing it
  (3) subject - my not doing it may be laid upon the account of my weariness

• verb type (Maekelberghe 2017; Fonteyn forthc.)
  (1) light - giving of thanks
  (2) state - the having of a sword
RESEARCH QUESTIONS

- Investigate relative importance of factors involved in the diachronic verbalization of the English gerund (NG > VG)

- Investigate the issue in aggregated vs. individual data:
  - Is there substantial inter-speaker variation, or, in other words, do we observe ‘individual conditioning’ of the existing variation?
  - If so, where/how does that individuality reveal itself in the individual’s linguistic behaviour?
CORPUS

Early Modern Multiloquent Authors (EMMA; Petré et al. 2018)
- Sample of 50 of the most prolific English writers born in the 17th century (mostly belonged to the London-based elite)
- 5 generations

In this study:
- 13 randomly selected speakers, born between 1600 and 1635
- Focus on prose and letters
- 23,681 ing-forms (including present participles)
- 6,074 nominal and verbal gerunds
## CORPUS

<table>
<thead>
<tr>
<th>Name</th>
<th>NG</th>
<th>VG</th>
<th>Name</th>
<th>NG</th>
<th>VG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heylyn, Peter</td>
<td>344</td>
<td>401</td>
<td>Dryden, John</td>
<td>90</td>
<td>320</td>
</tr>
<tr>
<td>Fuller, Thomas</td>
<td>172</td>
<td>271</td>
<td>Flavell, John</td>
<td>58</td>
<td>165</td>
</tr>
<tr>
<td>Milton, John</td>
<td>235</td>
<td>339</td>
<td>Fox, George</td>
<td>213</td>
<td>382</td>
</tr>
<tr>
<td>Taylor, Jeremy</td>
<td>102</td>
<td>380</td>
<td>Pierce, Thomas</td>
<td>91</td>
<td>295</td>
</tr>
<tr>
<td>Boyle, Robert</td>
<td>79</td>
<td>515</td>
<td>Swinnock, George</td>
<td>55</td>
<td>262</td>
</tr>
<tr>
<td>Boyle, Roger</td>
<td>79</td>
<td>192</td>
<td>Tillotson, John</td>
<td>82</td>
<td>176</td>
</tr>
<tr>
<td>Bunyan, John</td>
<td>401</td>
<td>374</td>
<td>Total</td>
<td>2001</td>
<td>4073</td>
</tr>
<tr>
<td>Name</td>
<td>NG</td>
<td>VG</td>
<td>Name</td>
<td>NG</td>
<td>VG</td>
</tr>
<tr>
<td>---------------------</td>
<td>------</td>
<td>------</td>
<td>---------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Heylyn, Peter</td>
<td>46.17%</td>
<td>53.83%</td>
<td>Dryden, John</td>
<td>21.95%</td>
<td>78.05%</td>
</tr>
<tr>
<td>Fuller, Thomas</td>
<td>38.83%</td>
<td>61.17%</td>
<td>Flavell, John</td>
<td>26.01%</td>
<td>73.99%</td>
</tr>
<tr>
<td>Milton, John</td>
<td>40.84%</td>
<td>59.16%</td>
<td>Fox, George</td>
<td>35.80%</td>
<td>64.20%</td>
</tr>
<tr>
<td>Taylor, Jeremy</td>
<td>21.16%</td>
<td>78.84%</td>
<td>Pierce, Thomas</td>
<td>23.58%</td>
<td>76.42%</td>
</tr>
<tr>
<td>Boyle, Robert</td>
<td>13.30%</td>
<td>86.70%</td>
<td>Swinnock, George</td>
<td>17.35%</td>
<td>82.64%</td>
</tr>
<tr>
<td>Boyle, Roger</td>
<td>29.15%</td>
<td>70.85%</td>
<td>Tillotson, John</td>
<td>31.78%</td>
<td>68.22%</td>
</tr>
<tr>
<td>Bunyan, John</td>
<td>51.74%</td>
<td>48.26%</td>
<td></td>
<td>31.6%</td>
<td>68.4%</td>
</tr>
</tbody>
</table>
METHODOLOGY

Decision tree classification models:
- can untangle the factors that contribute to a grammatical choice by showing which factors are more important (or ‘effective’) at different levels than others;
- And can be seen as an approximate reconstruction of the personal grammar of that individual (i.e. ‘typical’ co-selection of features).

- `rpart`: `gerund ~ det, func, verb type, genre, method = 'class', complexity for pruning = 0.01`
- No weighting of dependent variables applied (as to account for the frequency imbalance; the assumption is that there is an equal chance of NG and VG)
<table>
<thead>
<tr>
<th>Determiner</th>
<th>Function</th>
<th>Verb Type</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>BARE</td>
<td>BY, IN, FOR, OF, TEMP, ...</td>
<td>BY, IN, FOR, OF, TEMP, ...</td>
<td>PROSE</td>
</tr>
<tr>
<td></td>
<td>By ø destroying Souls, he ...</td>
<td>by onely torturing of men</td>
<td>LETTERS</td>
</tr>
<tr>
<td>POSS</td>
<td>his fearing God more then Man was ...</td>
<td>in the destroying of the ...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>his fearing God more then Man was ...</td>
<td>after his blaspheoming Shakespeare</td>
<td></td>
</tr>
<tr>
<td>THE</td>
<td>The seeing of our Friends in Heaven will ...</td>
<td>OBJECT</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>a cry will be among you, and a wishing you had never been born</td>
<td>SUBJECT COMPLEMENT</td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>... no reverencing of images</td>
<td>... that there should be christening of children</td>
<td></td>
</tr>
<tr>
<td>DEM</td>
<td>This forgetting of the God that saves us ...</td>
<td>SUBJECT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The laying down of life did abundantly proclaim his love</td>
<td>SUBJECT COMPLEMENT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>... whilst others make them groan, by abusing them to sin, and subjecting them to their lusts.</td>
<td>LEX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>whilst others make them groan, by abusing them to sin, and subjecting them to their lusts.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘LIGHT’</td>
<td>‘LIGHT’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>He is accus'd of Malevolence, and of taking Actions in the worst Sence</td>
<td>‘LIGHT’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>... that prayers, and supplication, and giving of thanks be made for all men</td>
<td>‘LIGHT’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>there is more required to make a good Scholler, then onely the having of many bookes</td>
<td>HAVE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HAVE</td>
<td></td>
</tr>
</tbody>
</table>
Decision trees

- **det = a, dem, quant, the**
  - yes -> VG (100%)
  - no
    - **verb_type = lex, light**
      - NG (29%)
        - NG (.82, .18)
      - VG (0%)
        - VG (.11, .89)
      - VG (71%)
        - VG (.12, .88)
TOP (BEFORE FIRST SPLIT)

- Predicted class
- Probability NG (left) vs. VG (right)
- Percentage of data set in node (at top node always 100%)
- Binary split based on (group of) features in variable that reduce entropy most successfully (a/dem/the vs. rest)

LOWER LEVEL

- Predicted class
- Probability NG (left) vs. VG (right)
- Percentage of data set in node
- Binary split based on (group of) features in variable that reduce entropy most successfully (lexical verbs vs. rest)

VG
- .13 .87
- 100%

yes det = a,dem,the no

NG
- .66 .34
- 16%

verb_type = lex
Decision trees

- Probability of VG in entire set
  - Suggested first split is determiner use
    - Further split into lexical verbs vs. light verbs and possessive have
      - If the base verb is **lexical** or **light** (29%), NG probability is .82
      - If the base verb is **have**, VG probability is .89
DECISION TREES
Decision trees
Decision trees
John Bunyan

- Probability of NG in entire set of JB is highest at .52
- Suggested first split is determiner use.
- Gerunds without determiner (51% of data) have probability of .75 to be VG.
- Further split into prep vs. non-prep contexts.
- If the bare gerund functions as a subject, sub comp, or object, it is most likely to be nominal (.84)
- Following a prep probability of VG is 0.79
Gerunds preceded by **a, no, or the** (70% of data) have probability of .55 to be VG.

Gerunds preceded by **possessive**, or Ø (70% of data) have probability of .95 to be VG.

Gerunds preceded by **no, a, or the** preceded by other prepositions or function as a subject are most likely to be NG if the verb is lexical, and VG if the verb is light or possessive have.

Gerunds preceded by **in**, **by**, or **pos** preceded by in, by, for, of or function as object have probability of .99 to be VG.
“lack of order or predictability; gradual decline into disorder”

<table>
<thead>
<tr>
<th>Boyle 1</th>
<th>Boyle 2</th>
<th>Bunyan</th>
<th>Dryden</th>
<th>Flavell</th>
<th>Fox</th>
<th>Fuller</th>
<th>Heylyn</th>
<th>Milton</th>
<th>Pierce</th>
<th>Swinnock</th>
<th>Taylor</th>
<th>Tillotson</th>
</tr>
</thead>
<tbody>
<tr>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
<td>det</td>
</tr>
<tr>
<td>verb</td>
<td>func, func</td>
<td>func, func</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>NA</td>
<td>verb</td>
<td>NA</td>
<td>NA</td>
<td>func, func</td>
<td>det, det, det</td>
<td>verb, func</td>
<td>det</td>
<td>NA</td>
<td>genre</td>
<td>func</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>det</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>det</td>
<td>det</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>func</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>func</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>det</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>genre</td>
<td>func</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>%</td>
<td>100%</td>
<td>71%</td>
<td>58%</td>
<td>33%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>entropy</td>
<td>0</td>
<td>1.29</td>
<td>1.38</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CONCLUSIONS

- The ‘relative nominality’ of the context has the highest ‘importance’ in the (individualised) decision trees, consistently returning as the factor that most ‘effectively’ explains NG vs. VG usage (even though the specificities are not always identical).
- But entropy increases at lower levels of tree:
  - This indicates, on the one hand, that individuality reveals itself only after a number of features have been selected.
  - If the particular factors employed to condition grammatical variation differ from individual to individual, the aggregate data set contains ‘contradictory’ information – which potential leads to ‘information loss’.
  - If more evidence is found for these personalised grammars (e.g. from forensic linguistic work and more research of this kind) then we must start treating the individual not as something to control but as something to study.
THANK YOU.

Lauren Fonteyn – l.fonteyn@hum.leidenuniv.nl
Andrea Nini – andrea.nini@manchester.ac.uk

University of Manchester
ICCG-10 - Sorbonne Paris Cité University (USPC) - July 16th-20th
REFERENCES