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

Lauren Fonteyn
Andrea Nini
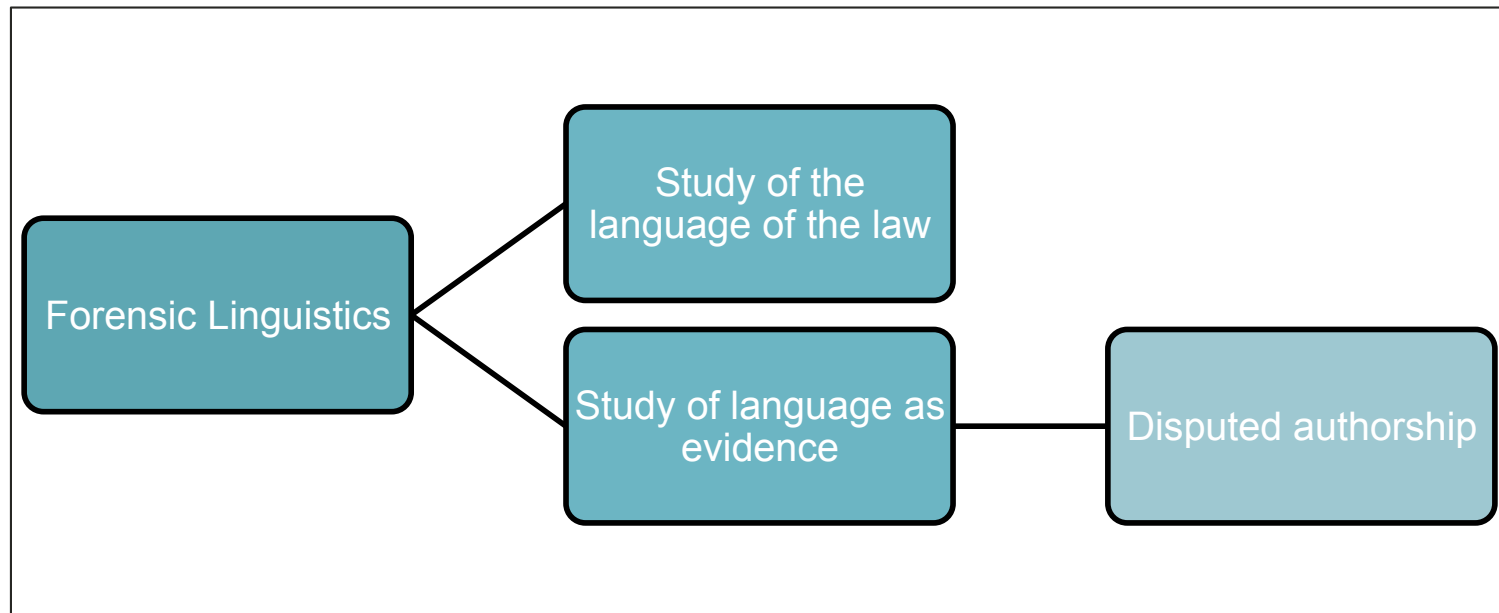
University of Manchester
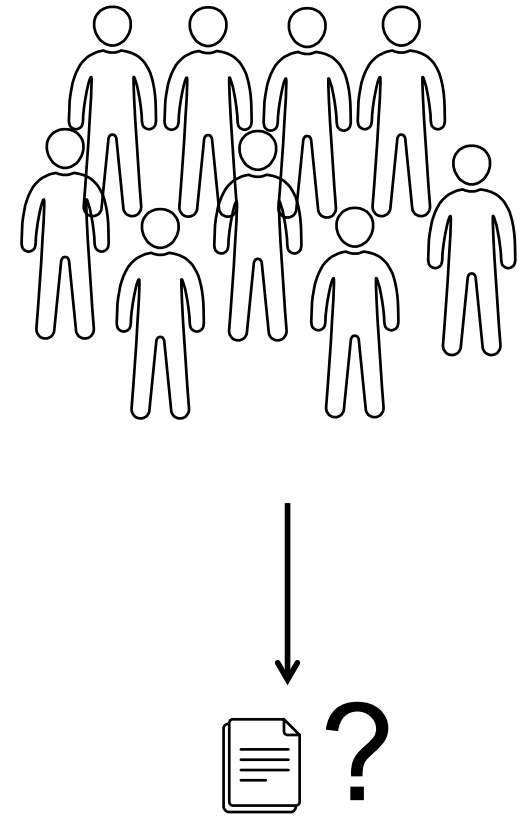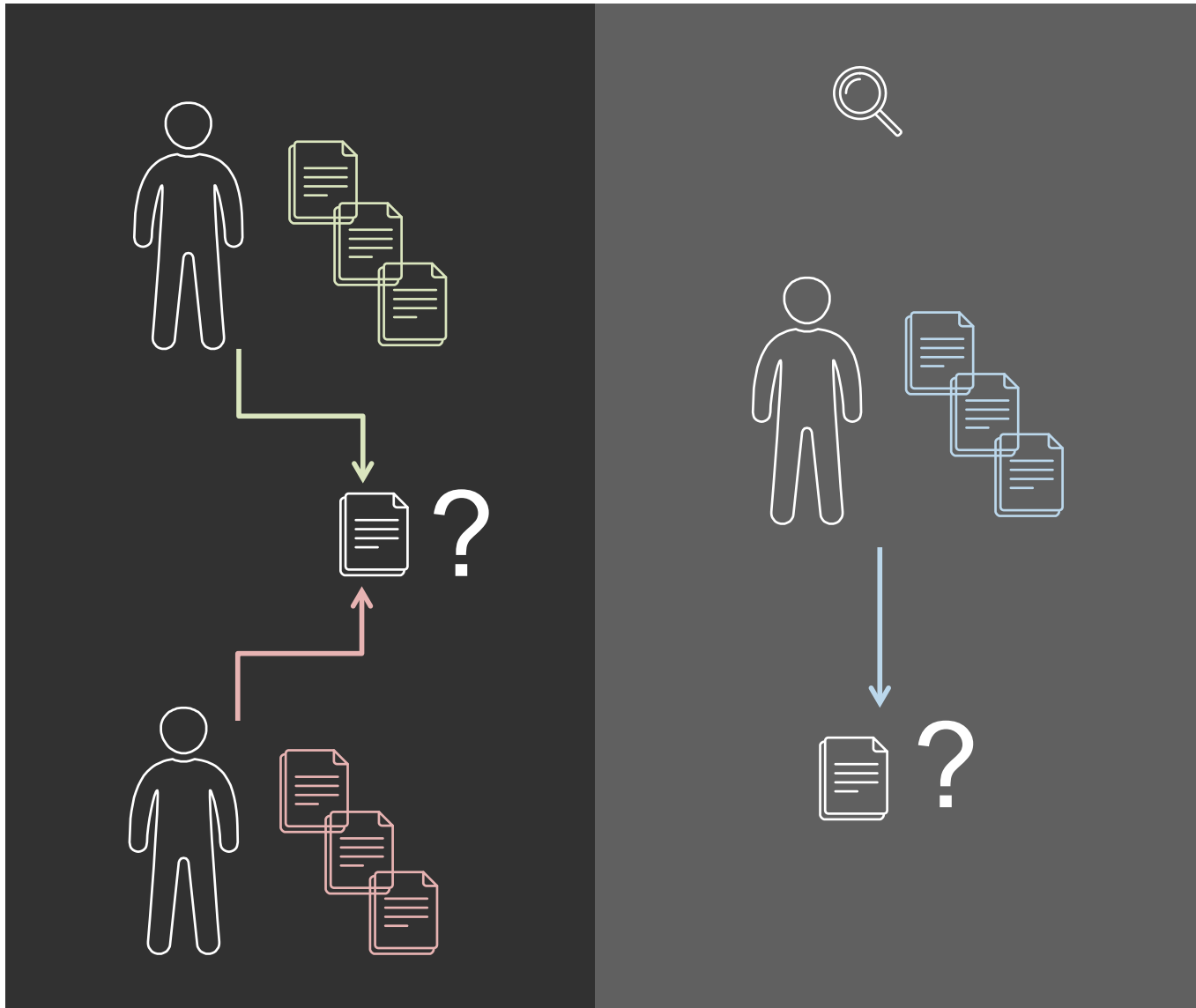
"

**The scientific study of language as applied to
forensic purposes and contexts**

(McMenamin 2002)

2

Forensic Linguistics

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

**AUTHOR A**

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 B**

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%

# Burrows (2002)

| Requirements | Findings | Explanation |
|---|---|---|
| • Most frequent 150 word types<br>• About 1000-1500 word tokens per text | • 85% of the times the real author is among the top 5 candidates | • "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:

$$\Delta_{(AB)} = \frac{1}{n} \sum \left| \frac{(A - \mu)}{\sigma} - \frac{(B - \mu)}{\sigma} \right|$$

"

# idiolect

*"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."*
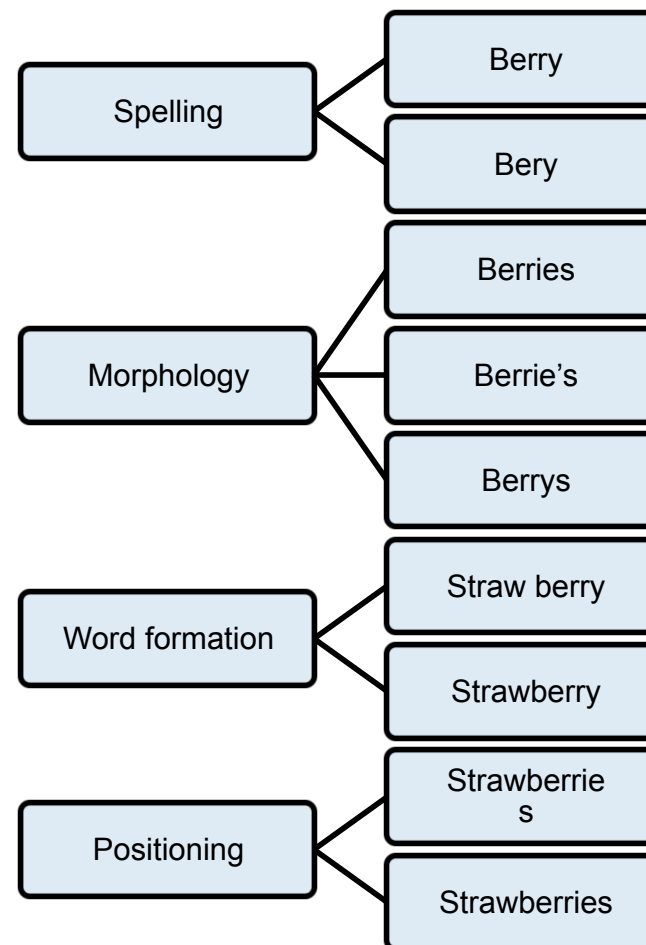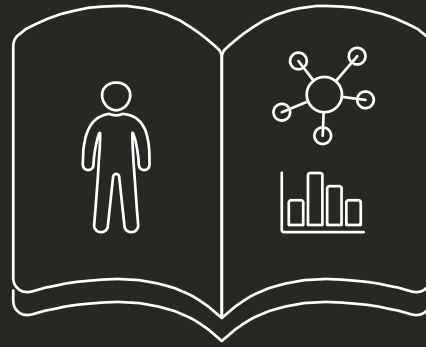
Bloch (1948)

**"**

# idiolectal co-selection

"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)

8

"STRAWBERRY"
Nos. 1-3

1. STRAWBERRIE
2. STRAWBERRI
3. STRWBERI

"STRAWBERRIES"
Nos. 4-23

4. STRAWBERRIES
5. STRAWBERRYS
6. STRAWBERRIE'S
7. STRAWBERRIS
8. STRAW BERRYS
9. STRAW-BERRIES
10. BoysenBerries
11. STRAWBERIES
12. STRAWBREEIE
13. STRAMBER IES
14a. STARWBERRY
14b. STARWBERRY
15. STRAWBRRY
16. STRAWBERRV
17a. STRAWBERR
17b. STRAWBERR
18a. STRAWBERRIE -S
18b. STRAWBERRIE -S
19. STRAWBERR -IES
20. STRAW- BERRYS
21. FRESH + S + [Image]
22. FRESH + [Images]
23. [Image]

McMenamin (2002) *Forensic Linguistics: Advances in Forensic Stylistics*

| | |
|---|---|
| Spelling | Berry |
| | Bery |
| Morphology | Berries |
| | Berrie's |
| | Berrys |
| Word formation | Straw berry |
| | Strawberry |
| Positioning | Strawberrie s |
| | 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*
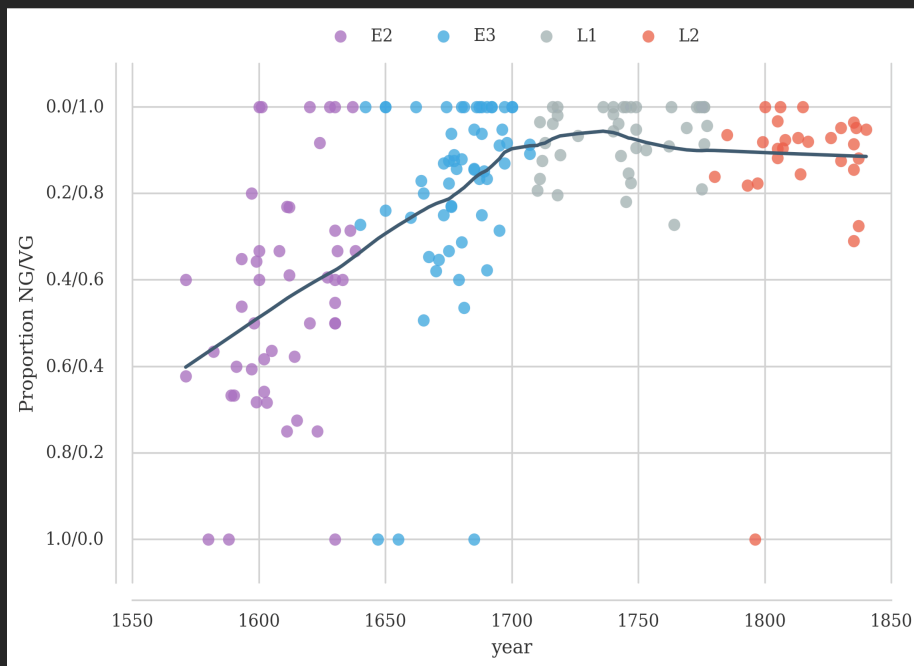(c. 1250 - e.g. *by writing a letter*)

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

## HISTORICAL DEVELOPMENT



(Figure taken from Fonteyn (2017) – data from PPCEME and PPCMBE)

- 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 'i*ng*-form network' – the functionally hybrid gerund splits into a nominal and a clausal component (Fonteyn forthc.).

> "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 and1635
- Focus on *prose* and *letters*
- 23,681 *ing*-forms (including present participles)
- 6,074 nominal and verbal gerunds

# CORPUS

| | NG | VG | | NG | VG |
|---|---|---|---|---|---|
| *Heylyn, Peter* | 344 | 401 | *Dryden, John* | 90 | 320 |
| *Fuller, Thomas* | 172 | 271 | *Flavell, John* | 58 | 165 |
| *Milton, John* | 235 | 339 | *Fox, George* | 213 | 382 |
| *Taylor, Jeremy* | 102 | 380 | *Pierce, Thomas* | 91 | 295 |
| *Boyle, Robert* | 79 | 515 | *Swinnock, George* | 55 | 262 |
| *Boyle, Roger* | 79 | 192 | *Tillotson, John* | 82 | 176 |
| *Bunyan, John* | 401 | 374 | *Total* | **2001** | **4073** |

# CORPUS

| | NG | VG | | NG | VG |
|---|---|---|---|---|---|
| *Heylyn, Peter* | 46.17% | 53.83% | Dryden, John | 21.95% | 78.05% |
| *Fuller, Thomas* | 38.83% | 61.17% | Flavell, John | 26.01% | 73.99% |
| *Milton, John* | 40.84% | 59.16% | Fox, George | 35.80% | 64.20% |
| *Taylor, Jeremy* | 21.16% | 78.84% | Pierce, Thomas | 23.58% | 76.42% |
| *Boyle, Robert* | 13.30% | 86.70% | Swinnock, George | 17.35% | 82.64% |
| *Boyle, Roger* | 29.15% | 70.85% | Tillotson, John | 31.78% | 68.22% |
| *Bunyan, John* | 51.74% | 48.26% | | **31.6%** | **68.4%** |

# 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)

| Determiner | Function | Verb Type | Genre |
|---|---|---|---|
| **BARE**<br>▫ *By ø destroying Souls*, he …<br><br>**POSS**<br>▫ *his fearing God more then Man* was …<br><br>**THE**<br>▫ *The seeing of our Friends in Heaven* will ...<br><br>**A**<br>▫ *a cry* will be among you, and *a wishing you had never been born*<br><br>**NO**<br>▫ … **no** reverencing of images<br><br>**DEM**<br>▫ *This forgetting of the God that saves us* … | **BY, IN, FOR, OF, TEMP, …**<br>▫ **by** *onely torturing of men*<br>▫ **in** *the destroying of the* …<br>▫ **aft**er *his blaspheming Shakespeare.*<br><br>**OBJECT**<br>▫ I would seriously **recommend** *the Arming of our Pikemen*<br><br>**SUBJECT COMPLEMENT**<br>▫ … that **there** should **be** christening of children<br>▫ It is not the giving out of mercy that troubles him, but …<br><br>**SUBJECT**<br>▫ The laying down of life **did** abundantly proclaim his love | **LEX**<br>▫ … whilst others make them groan, by **abusing** them to sin, and **subjecting** them to their lusts.<br><br>**'LIGHT'**<br>▫ He is accus'd of Malevolence, and of **taking** Actions in the worst Sence<br>▫ … that prayers, and supplication, and **giving** of thanks be made for all men<br><br>**HAVE**<br>▫ there is more required to make a good Scholler, then onely the **having** of many bookes | **PROSE**<br><br><br>**LETTERS** |

# Decision trees
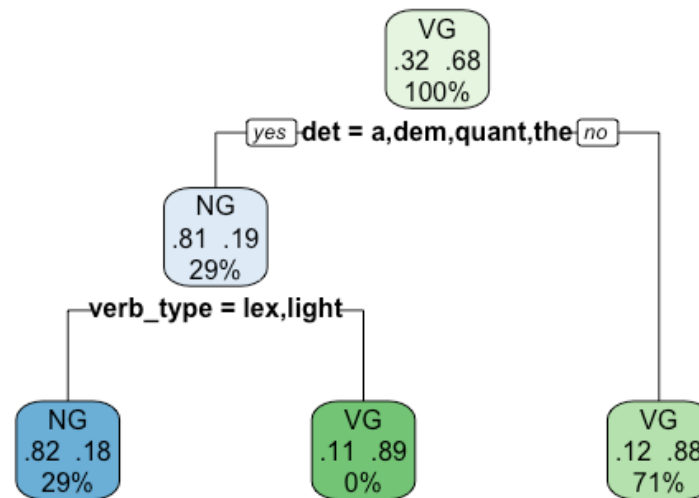
## TOP (BEFORE FIRST SPLIT)

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)
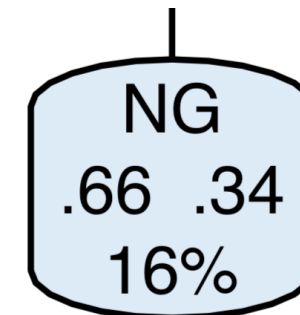
VG
.13  .87
100%
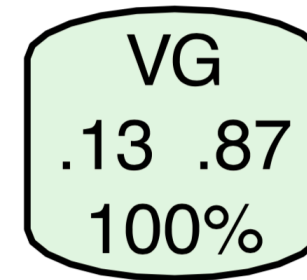
yes — det = a,dem,the — no

## LOWER LEVEL

predicted class

probability NG (left) vs. VG (right)
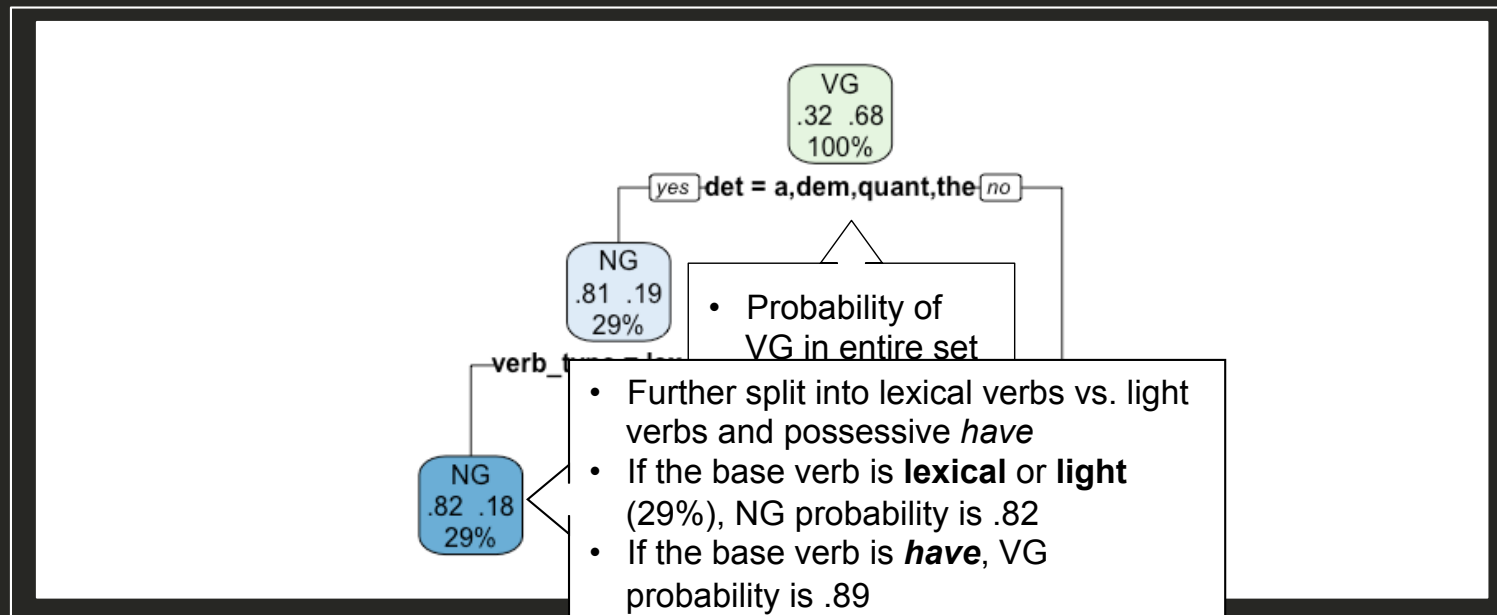
percentage of data set in node

binary split based on (group of) features
in variable that reduce entropy most
successfully (lexical verbs vs. rest)

NG
.66  .34
16%

verb_type = lex

# Decision trees

VG
.32  .68
100%

yes — **det = a,dem,quant,the** — no

NG
.81  .19
29%

- Probability of
  VG in entire set

verb_type = lex

NG
.82  .18
29%

- 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
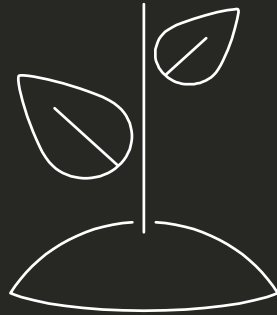
DECISION TREES
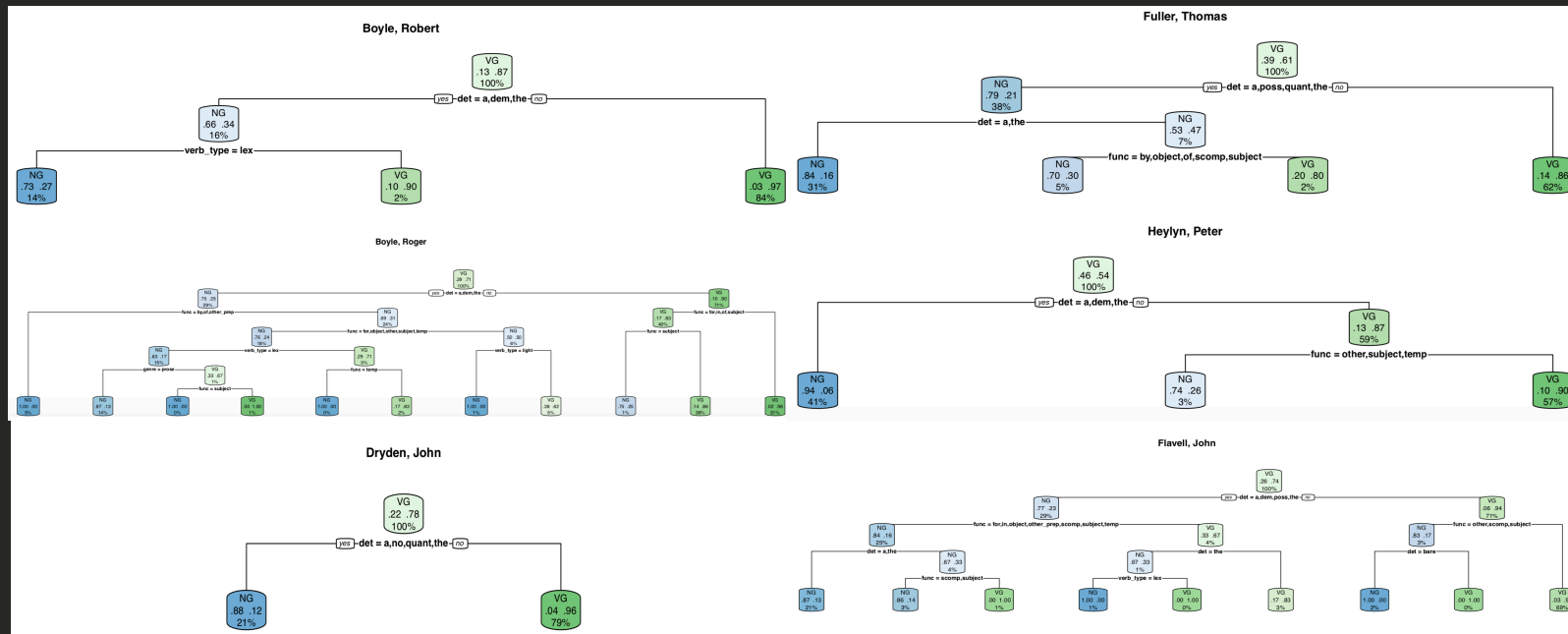
# Decision trees

# Decision trees

# John Bunyan



NG
.52 .48
100%

- 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.

VG
.25 .75
51%

object,scomp,subject

NG
.80 .20
49%

NG
.84 .16
3%

VG
.21 .79
48%

- 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
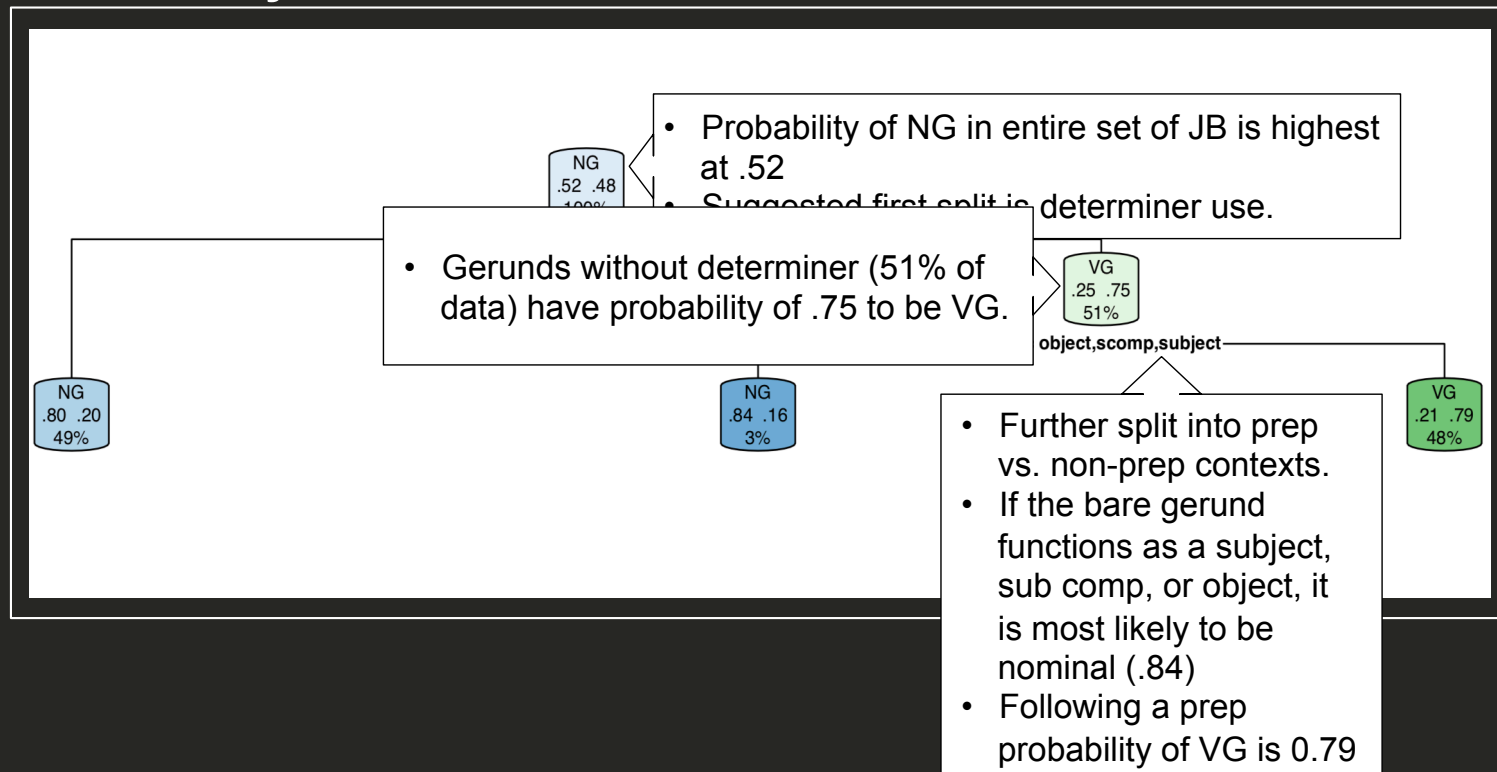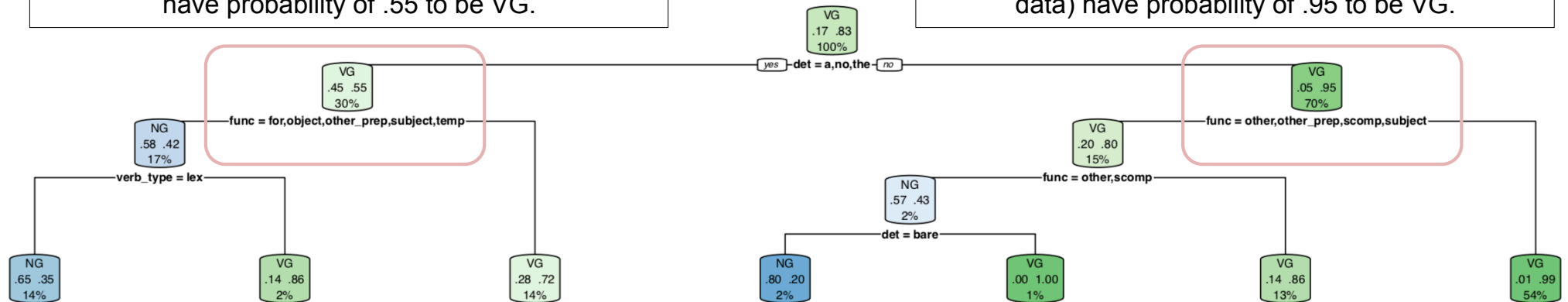
# George Swinnock



Swinnock, George

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.*

e preceded by *in,* 72 to be VG.

Gerunds preceded by **poss** or Ø preceded by *in, by, for, of* or function as object have probability of .99 to be VG.

28

# John Dryden
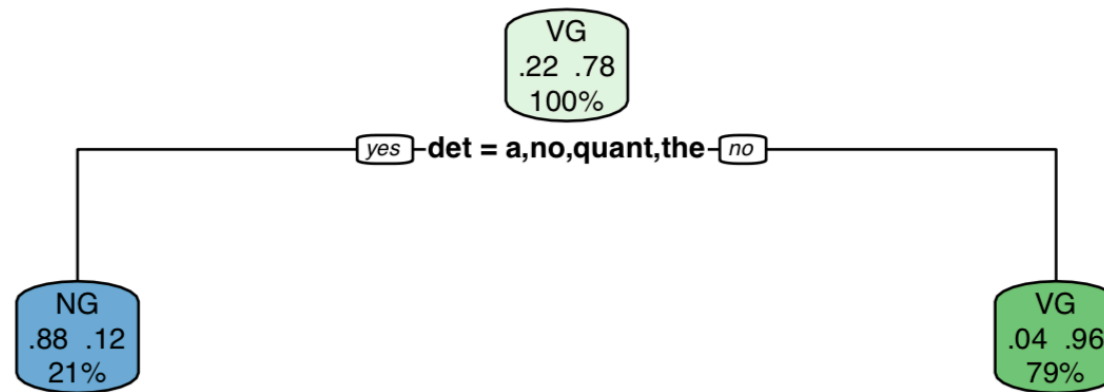
## ENTROPY

"lack of order or predictability; gradual decline into disorder"

| | Split1 | Split2 | Split3 | Split4 |
|---|---|---|---|---|
| *Boyle 1* | det | verb | NA | NA |
| *Boyle 2* | det | func, func | func, func | verb |
| *Bunyan* | det | func | NA | NA |
| *Dryden* | det | NA | NA | NA |
| *Flavell* | det | func, func | det, det, det | verb, func |
| *Fox* | det | NA | NA | NA |
| *Fuller* | det | det | func | NA |
| *Heylyn* | det | func | NA | NA |
| *Milton* | det | NA | NA | NA |
| *Pierce* | det | genre | func | NA |
| *Swinnock* | det | func, func | verb, func | det |
| *Taylor* | det | func, func | func, func | det |
| *Tillotson* | det | det | verb | func |
| **%** | **100%** | **71%** | **58%** | **33%** |
| **entropy** | **0** | **1.29** | **1.38** | **1.58** |

## 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

Coulthard, M. 2004. Author identification, idiolect, and linguistic uniqueness. Applied Linguistics 25. 431–447.

Croft, W. & D. A. Cruse. 2004. Cognitive linguistics. Cambridge: Cambridge University Press.

De Smet, H. 2013. Spreading patterns: Diffusional change in the English system of complementation. Oxford: Oxford University Press.

Fanego, T. 2004. On reanalysis and actualization in syntactic change: The rise and development of English verbal gerund. Diachronica 21(1). 5–55.

Hoffmann, T. 2017. Construction grammars. In B. Dancyger (ed.), The Cambridge handbook of cognitive linguistics. Cambridge: Cambridge University Press.

Lass, R. 1992. Phonology and Morphology. In N. Blake (ed.), The Cambridge history of the English language. Vol 2. 1066-1476, 23–154. Cambridge: Cambridge University Press.

Killie, K. & T. Swan. 2009. The grammaticalization and subjectification of adverbial -ing clauses (converb clauses) in English. English Language and Linguistics 13. 337–363.

Petré, P., L. Anthonissen, S. Budts, E. Manjavacas, W. Standing, E.-L. Silva & O. Strik. 2017. Early Modern Multiloquent Authors (EMMA). Antwerp: Linguistics Department.

van de Pol, N. & P. Petré. 2015. Why is there a Present-Day English absolute? English Studies 39(1). 199–229.