Lauren Fonteyn* and Andrea Nini

Individuality in syntactic variation: An investigation of the seventeenth-century gerund alternation

https://doi.org/10.1515/cog-2019-0040
Received 06 May 2019; revised 15 October 2019; accepted 04 February 2020

Abstract: This study investigates the extent to which there is individuality in how structural variation is conditioned over time. Earlier research already classified the diachronically unstable gerund variation as involving a high fraction of mixed-usage speakers throughout the change, whereby the proportion of the conservative variant versus the progressive variant as observable in the linguistic output of individual language users superficially resembles the mean proportion as observable at the population level. However, this study sets out to show that there can still be heterogeneity within such a centralized population in terms of how each individual conditions the observed variation. A random forest and conditional inference tree analysis of over 14,000 gerunds uttered by nineteen seventeenth-century authors is presented to show that, while the most important language-internal factors conditioning the gerund variation are adopted by (and shared between) all authors, we can still attest inter-individual variation (i) at lower levels of variable importance, and (ii) in the breadth of the range of contexts individual authors employ to condition the attested variation.

Keywords: gerund, idiolect, random forest, conditional inference tree, usage-based

1 Introduction

While there are certainly good reasons to use aggregate, population-level data to study language variation and change (Eckert 2019), it must be acknowledged that this 'aggregate approach' also comes with certain problems and pitfalls, as studies that solely focus on population-level change are not designed to address

***Corresponding author: Lauren Fonteyn,** Leiden University Centre for Linguistics (LUCL), Leiden University, Leiden, 2300 RA Netherlands, E-mail: l.fonteyn@hum.leidenuniv.nl
Andrea Nini, Department of Linguistics and English Language, University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland, E-mail: andrea.nini@manchester.ac.uk

some important open questions in the study of language change. Moreover, the general lack of attention devoted to the degree of divergence in the linguistic behaviour of (historical) individuals has also left us somewhat "ignorant" in our understanding of variation at the level of the individual (MacKenzie 2019), which is a crucial component of more applied branches of linguistic analysis, such as literary or forensic authorship analysis (Burrows 2002: 282; Coulthard 2004).

However, in recent years, a growing group of historical (socio-)linguists has placed more explicit emphasis on using corpus data to study individual language use (e. g., papers in this issue; Schmid and Mantlik 2015; Feltgen et al. 2017; Hundt et al. 2017; Petré 2017; Petré and Van de Velde 2018). In two notable studies, Nevalainen et al. (2011) and Baxter and Croft (2016) set out to investigate how individuals behave with respect to morphological and syntactic variation that is diachronically unstable at the population level. Using historical corpus data, Nevalainen et al. document how individual language users 'participate' in six different changes as they unfolded over the course of the Early Modern English period, including morpheme replacement (e. g., *ye* for second person plural was replaced by *you*) and changes in more abstract structural patterns (e. g., negation). From their study, Nevalainen et al. conclude that, when confronted with the fact that there are "different ways to say the same thing" (Labov 1972: 188), individual language users can either consistently opt for one of the different variants at all times, or they can participate in the variability that is observable in aggregate, population-level language, by actually using both variants in alternation. They furthermore add that the *degree* to which individual language users participate in the variability visible at the population level can be predicted by (i) the type of linguistic change, (ii) the rate of diffusion of the process over time, and (iii) the stage of development the change is in. More specifically, their results indicate that the fraction of individuals that actually engage in variable or 'mixed' usage is highest with ongoing changes in abstract structural patterns, which are generally more protracted than simple morpheme replacement processes. Baxter and Croft (2016) subsequently revisit the data set examined in Nevalainen et al. (2011), concluding that there is indeed a correlation between the fraction of mixed-usage and the overall rate of diffusion of the new variable. This correlation, they argue, can be explained by "the interaction of the differential weighting of the variants and the degree of accommodation of the speakers" (Labov 1972: 188).

It is difficult to overestimate the value of these accounts for understanding the relation between the type, stage of development, and pace of linguistic change on the one hand, and the extent to which individuals actually engage in mixed usage (throughout their lifespan) on the other. Yet, an issue that is not explicitly dealt with in these analyses is how such mixed-usage individuals condition the attested grammatical variation, and whether they behave homogeneously in that respect.

Such homogeneity – or rather, 'orderly heterogeneity' (Weinreich et al. 1968) – is taken for granted: variationist research consistently seems to point out that the constraints or conditions in which the variants are used are *shared* by all individuals in a population or community (MacKenzie 2019: 4; in reference to: Poplack and Tagliamonte 1991; Meyerhoff and Walker 2007; Labov 2012: 265; Tagliamonte 2013). Notably, this is considered to be the case even despite fluctuations in the rate by which individual speakers use one or the other variant.

However, there are non-trivial reasons to believe that, even when a given population consists predominantly (or even entirely) of mixed-usage speakers, this does not necessarily mean the population is entirely or even largely homogeneous in its linguistic behaviour. In diachronic Construction Grammar, diachronic change is often seen as proceeding gradually from context to context, with new structures becoming possible in some contexts before others (Traugott and Trousdale 2013; 'constructional diffusion' in Rutten and van der Wal 2014; Smirnova 2015). As such, different individuals that engage in mixed-usage of old and new variants may still behave fully progressively or conservatively with respect to the precise contexts or conditions in which they find old or new variants appropriate. As a result, there may still be substantial heterogeneity in the linguistic behaviour of what appears to be a homogeneous mixed-usage population: given that each individual speaker is a product of a unique social history (e. g., Johnstone 1996; Eckert 2019), and given that they all receive different inputs that they may process differently (see Dąbrowska Forthcoming), it is also possible that each individual has acquired a different system to condition the use of equivalent structures. This possibility is strongly implied in usage-based theories of language, which model language as a complex adaptive system resulting from individuals' encounters with exemplars (Beckner et al. 2009). As such, usage-based theory indirectly predicts that each individual will be unique in their personal model of their language – that is, their own *idiolect*.

Although the work on idiolect is still in its infancy, there is increasing evidence of the validity of this prediction. For example, Mollin (2009) and Barlow (2013) investigated the idiolect of UK and US politicians, revealing that the particular combinations of words or parts of speech they used were consistently idiosyncratic over time. In writing, it has been shown that even relatively formulaic registers such as business emails can have a large component of individuality in word combinations (Johnson and Wright 2014; Wright 2017). In terms of phonological and grammatical variability, MacKenzie (2019: 2) Mackenzie (2019: 2) argues that "individuals can differ in their mental representation of a particular surface structure", and Dąbrowska (2012; 2018) has found experimental evidence that not all native speakers of a language have precisely the same understanding of even the most common grammatical

constructions, such as the passive voice. Furthermore, in diachrony, Schmid and Mantlik (2015) studied the development of the construction [noun BE *that*] (e. g., *the fact is that* ...) and discovered how individuality played a role in the evolution of this construction. Finally, major indirect evidence of a substantial element of individuality in language comes from the work in the field of authorship analysis, both for forensic and literary purposes. Thus far, most of the work in this area has shown that the author of an anonymous document can be discovered in a pool of candidates with relatively high degrees of accuracy depending on the amount of data available by simply using frequencies of words or character sequences (Stamatatos 2009; Grieve 2007). Although these techniques are not transparent in revealing the mechanisms underlying individuality in linguistic production, the fact that it is possible to attribute texts to their authors suggests that the amount of individuality in language is not trivial, and thus it challenges the notion that "individual variation is reduced below the level of linguistic significance" (Labov 2012: 265).

The aim of this paper, then, is to investigate the extent and nature of individuality in how competing variable structures are functionally constrained or conditioned. To this end, we focus on the variable use of two different types of gerunds (i. e., deverbal nominalizations formed with the *-ing* suffix) that were competing over the same (grammatical) contexts in the seventeenth century, using two types of multifactorial models that have recently started to gain traction in variationist studies: random forests, and conditional inference trees (Tagliamonte and Baayen 2012; other applications in, e. g., Robinson 2011; Kerz and Wiechmann 2013; Claes 2016; Szmrecsanyi et al. 2016). By means of these models, we were able to demonstrate that, while some evident constraints underlying the use of grammatical variants are indeed shared, there still appear to be some subtle, yet non-trivial differences between individuals.

2 The gerund alternation

As pointed out by Nevalainen et al. (2011), the relation between individual usage and the population mean differs between different types of linguistic variation. In principle, a 50–50 distribution of older (conservative) and newer (progressive) variants at the population level can emerge from two types of patterns: the population statistics may either arise from a highly polarized pattern, where 50% of the speakers in the population are consistently conservative or progressive, or we may also observe a centralized population, where all speakers actually participate in mixed usage (cf. Figure 1).

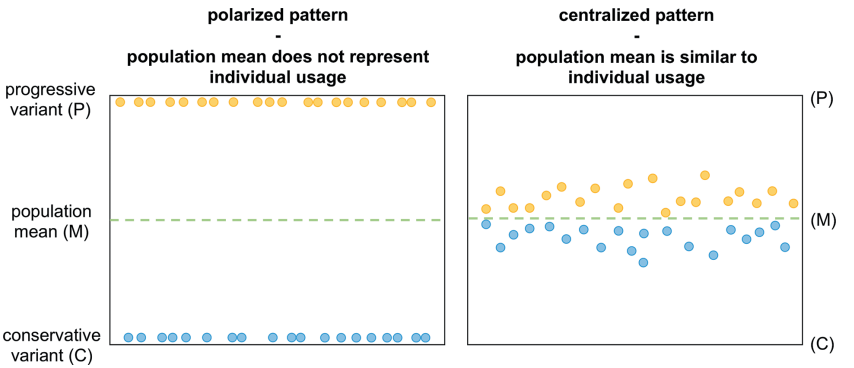

Figure 1: Patterns of individual usage, as compared to a 50–50 distribution of progressive and conservative variants at the population mean.

Having examined such patterns for a number of morpho-syntactic variation pairs, Nevalainen et al. (2011) found that a more centralized pattern tends to be more common when the competing variants are (abstract) syntactic constructions. The most centralized pattern with highest fraction of actual variable users was attested with what they termed "the (ing) variable", illustrated in example (1) to (4).

(1) ... as I was betwixt Elstow and Bedford, the temptation was hot upon me to try if I had Faith by *doing **of** some miracle*; (EMMA, John Bunyan, 1666)

(2) ... and revers'd the Laws, establish'd in the Universe, to engage Men to obey those of God, By *doing miracles* so numerous and great (EMMA, Robert Boyle, 1659)

(3) So that *the leaving **of** the word, Altar, out of the Common Prayer booke last established*, and other altera_ions which ... (EMMA, Peter Heylyn, 1636)

(4) ... yet the comfort of them extends not to the former, for want of this propriety in them. The *leaving out one word in a Will*, may marr the estate and disapoint all a mans hopes; (EMMA, George Swinnock, 1662)

The constructions exemplified in (1)-(4) – often referred to as 'gerunds' – share a number of characteristics. The most obvious similarity between the constructions is that they are all formed by means of the derivational suffix *-ing* attaching to a verb (that is, *do* in example (1)-(2) and *leave out* in example (3)-(4)). The resulting structure takes largely the same distribution as a noun phrase, serving as a subject (e. g., in examples (3)-(4)), direct object, or prepositional complement (e. g., in examples (1)-(2)) in a larger clause. The constructions vary, however, in their internal structure: in example (1) and (3), the internal structure of the gerund closely resembles that of a typical English noun phrase, with the 'object' of *do* and *leave out* occurring as an *of*-phrase. By contrast, the gerund in (2) and (4), does not require an *of*-phrase.

Over the past few decades, a number of studies have been devoted to charting and explaining the emergence and historical development of the latter type of gerund. In a nutshell, it is clear that the two structural manifestations of the gerund are historically related, with the periphrastic *of*-phrase structure being the 'older' variant (Fonteyn 2019). As explained by Fanego (2004) and De Smet (2008, 2013), the first examples of the non-periphrastic (i. e., '*of*-less') variant emerged in the Middle English period in very specific contexts. Over the course of the Middle and Modern English period, then, this new variant further diffused to different grammatical contexts (Fanego 2004).

Given that this diffusion seems to involve a high fraction of mixed-usage speakers throughout the change (Nevalainen et al. 2011), it is tempting to conclude that slow, gradual changes of abstract syntactic patterns such as the ones affecting the English gerund are essentially changes whereby the behaviour of individual language users aligns with the trends we observe in aggregate language. However, it is important to point out that such conclusions cannot and should not be drawn, as the study by Nevalainen et al. (2011) only considers the overall proportions of the old variants versus its newer competitor for each individual in a larger population. Thus, the question whether (and how) different individuals condition variation is not explicitly treated. This is important because it is not necessarily the case that, for example, every 50–50 distribution emerges from the same underlying conditions. Consider, for instance, two individuals who eat four meals a day: the first individual, Frank, eats a hot breakfast every day, and a hot snack in the afternoon, while his lunches and dinners are cold meals. Now consider another individual, Laura, whose eating habits also involve a 50–50 distribution of hot and cold meals. However, Laura eats a cold meal for breakfast, another cold meal in the afternoon, and a hot dish for lunch and dinner, which means that the times of the day – or, the conditions – on which Frank and Laura eat these hot and cold meals are entirely different (cf. Figure 2).

It is not unthinkable that a similar situation applies to cases of grammatical variation. In particular, it may apply in cases where the diachronic outcome of competition between variants is not a gradual replacement (as is the case for the majority of the variation pairs examined by Nevalainen et al. (2011) and Baxter and Croft (2016)), but a more stable and *conditioned* co-existence of the variant forms. This appears to be the case for the structural gerund variation: as briefly noted by Baxter and Croft (2016: 163) and Nevalainen et al. (2011: 7), the relative frequency of the gerund with direct object as compared to its periphrastic counterpart seems to stabilize around 80%. One explanation that has been offered for the 'survival' of periphrastic gerunds suggests that the initially

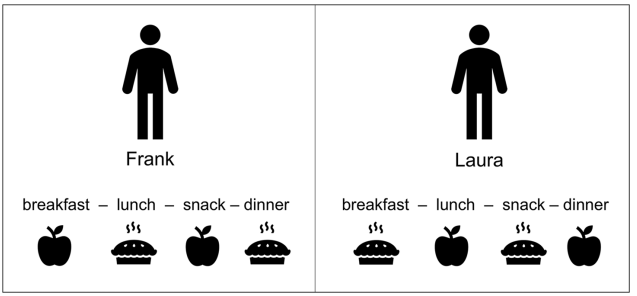

Figure 2: Illustration of deviating underlying conditions leading to the same superficial distribution of two variants.

unconditioned alternation pair gradually became conditioned: the new variant gradually came to replace its predecessor in a large number of contexts, while the older variant specialized to a (largely) distinct set of functions (Fonteyn 2019). Thus, rather than a complete replacement, the process is perhaps best characterized as one where the functional competition between the two formal variants very slowly and gradually led to a re-negotiated 'peace treaty' between them (see, e. g., Fonteyn (2017, 2019); for a functional-semantic analysis showing in Present-day English there is no longer a true 'gerund alternation', see Maekelberghe (2017)).

Thus, given the complexity of this development, it would be far too coarse-grained to examine the homo- or heterogeneity of a population of language users based on their general usage-rates of new and old variants alone; it is only by homing in on the individual, and understanding the conditioning of the variants at that level of granularity, that further complementary insights can be gained. To do so, as shown in the remainder of this paper, we can use a statistical toolkit containing two multifactorial models.

3 Data and methodology

3.1 Corpus, data and method

To investigate individuality in gerund usage, this study relies on data drawn from the corpus of *Early Modern Multiloquent Authors* (EMMA), which is a "large-scale specialized corpus [that] allows us to explore the interaction between the

individual and aggregate levels" (Petré et al. 2019: 83). Compared to other historical corpora, the size of the entire corpus is vast (90,000,000 words), making it particularly suitable for any kind of quantitative and statistical analysis.

The present study focuses on writings classified as 'prose' or 'letters' by authors from the first three generations in the corpus. Within each generation, authors were selected at random, the sole criteria being that the author produced writings in both genres under scrutiny. The number of authors per generation is determined by the frequency of the least frequent outcome variable (which should exceed a minimum of 1,000 tokens). For each of these authors, the total number of gerunds was collected. Because of the vast size of EMMA, a substantial number of tokens was found for each author, the smallest and largest number of observations being 223 for John Flavell and 2,457 for Gilbert Burnet respectively.

In line with Nevalainen et al. (2011: 13), we consider the analysis of the gerund alternation to be "a binary analysis of the use of *of*-phrases (...) vs. direct objects". As such, we distinguished two variants:

- ing-OF: *ing*-nominals that take an *of*-phrase (e. g., *eating of meat*)
- ing-Ø: *ing*-nominals that take a direct object (e. g., *eating meat*)

Gerunds that did not have explicit objects (e. g., *drinking*), or where the *of*-phrase expressed the subject of the *ing*-form (e. g., *galloping of horses*) were discarded. The resulting number of ing-OF and ing-Ø tokens, as well as their relative frequency for each author and each generation, is listed in Table 1. In total, 14,078 gerunds were collected for further statistical analysis.

As indicated in Section 1, the present analysis deviates from earlier studies such as Nevalainen et al. (2011) in that it does not solely offer a comparative picture of the relative frequency of two variables as used by each individual author. Rather, we consider the variation between ing-OF and ing-Ø as the *dependent variable* in a multifactorial analysis. In other words, it will be investigated whether the 'choice' between the two variants can be explained by a number of factors or *independent variables* (see Section 2.2–2.3). More specifically, the data will be analysed by means of two types of classification or 'decision tree' based methods (implemented in R): the conditional inference tree (ctree{party}) and random forest algorithm (cforest{party}). Scripts and data are available under a Creative Commons license and can be retrieved from: http://doi.org/10.5281/zenodo.3692868.

Both methods operate on a similar principle: confronted with a dataset in which tokens can be coded as either ing-OF or ing-Ø, the models seek to predict whether it is more likely to encounter ing-OF or ing-Ø given a number of predictor categories along which the data set can be partitioned. In doing so, the algorithm *recursively* works to split the dataset into smaller sets in which

Table 1: Absolute and relative frequencies of ing-OF and ing-Ø per author and per generation in EMMA.

Generation 1 (1599–1613)	Generation 2 (1621–1627)	Generation 3 (1639–1644)
101. Heylyn, Peter (1599–1662)/**HP1** OF: 344 (46.17%) Ø: 401 (53.83%)	201. Boyle, Roger (1621–1679)/**BRG2** OF: 79 (29.15%) Ø: 192 (70.85%)	305. Mather, Increase (1639–1723)/**MI3** OF: 201 (23.93%) Ø: 639 (76.07%)
102. Prynne, William (1600–1669)/**PW1** OF: 496 (47.06%) Ø: 558 (53.83%)	202. Pierce, Thomas (1622–1691)/**PT2** OF: 91 (23.58%) Ø: 295 (76.42%)	308. Crouch, Nathaniel (1640–1725)/**CN3** OF: 197 (19.64%) Ø: 806 (80.36%)
104. Fuller, Thomas (1607–1661)/**FT1** OF: 172 (38.83%) Ø: 271 (61.17%)	204. Fox, George (1624–1691)/**FG2** OF: 213 (35.80%) Ø: 382 (64.20%)	310. Behn, Aphra (1640–1689)/**BA3** OF: 18 (6.57%) Ø: 256 (93.43%)
105. Milton, John (1608–1674)/**MJ1** OF: 235 (40.94%) Ø: 339 (59.06%)	205. Boyle, Robert (1627–1691)/**BRB2** OF: 79 (13.30%) Ø: 515 (86.70%)	312. Burnet, Gilbert (1643–1715)/**BG3** OF: 509 (20.72%) Ø: 1948 (79.28%)
106. Taylor, Jeremy (1613–1667)/**TJ1** OF: 102 (21.16%) Ø: 380 (78.84%)	206. Swinnock, George (1627–1673)/**SG2** OF: 55 (17.35%) Ø: 262 (82.65%)	314. Penn, William (1644–1718)/**PW3** OF: 571 (24.02%) Ø: 1806 (75.98%)
	207. Bunyan, John (1628–1688)/**BJ2** OF: 401 (51.74%) Ø: 374 (48.26%)	
	208. Flavell, John (1630–1691)/**FJ2** OF: 58 (26.01%) Ø: 165 (73.99%)	**Total Generation 1/G1** OF: 1349 (40.90%) Ø: 1949 (59.10%)
	209. Tillotson, John (1630–1694)/**TJ2** OF: 82 (31.78%) Ø: 176 (68.22%)	**Total Generation 2/G2** OF: 1148 (29.98%) Ø: 2681 (70.02%)
	210. Dryden, John (1631–1700)/**DJ2** OF: 90 (21.95%) Ø: 320 (78.05%)	**Total Generation 3/G3** OF: 1496 (21.52%) Ø: 5455 (78.84%)

variation is optimally reduced. The resulting picture is a recursive (binary) partitioning that resembles a hierarchical tree, where the order of predictor selection reflects the reduction in uncertainty about the dependent variable

after seeing a predictor (in ctree and cforest splits are selected based on conditional inference tests (Hothorn et al. 2006)). With the conditional inference tree algorithm, a single tree is produced, whereas with the random forest algorithm, multiple inference trees with randomly sampled feature sets are combined into a 'forest' (Breiman 2001; Hothorn et al. 2006; Strobl et al. 2009), which is a form of bootstrapping that yields a more robust ranking of the importance of predictors, as to counter over-fitting and lead to more generalisable results. In the present study, the number of trees (ntree) is set to 1,000.

The application of random forests – and, to a lesser extent, conditional inference trees – to address (socio-)linguistic research questions is relatively new. Still, they should be considered important additions to the (historical) socio-linguist's methodological toolkit, "because random forests work with samples of the predictors, [and therefore] they are especially well applicable to problems with more variables than observations" (Tagliamonte and Baayen 2012: 24; Janitza et al. 2013). As such, the method alleviates a potential problem in historical linguistics, where researchers interested in quantitative and statistical analysis can be confronted with a small number of observations (because of limited data), but have an extensive set of predictors. Relatedly, random forest models have no problems with data sparsity or perfect separation of response classes, which means there is no need to control the number of (and/or levels of) predictors to be considered in the model. Moreover, as illustrated by Tagliamonte and Baayen (2012), the individual language users in the sample can be included in the model as predictors to quantify the extent to which idiolect helps predict the dependent variable. In their own case study, Tagliamonte and Baayen (2012: 25) found that the variation between *was* and *were* in York English was most effectively reduced if the data set was first split by sets of individuals, rather than by gender, level of education, age of the speakers, or any language-internal predictors (Tagliamonte and Baayen 2012: 25). Such a result is indicative of a noteworthy degree of individuality with regard to how variation is grammatically conditioned. In other words, while variation in language may be systematically conditioned by means of social and grammatical conditions, it is possible that individuals do not apply the entire range of conditions to the same extent.

To investigate how individuals behave with respect to language internal factors to condition variation, it is interesting to model the data by means of a single conditional inference tree as well. Random forests are often discussed in combination with conditional inference trees. The reason for this is that random forests are superior for statistical prediction and classification, but are very difficult to interpret, as they provide no insight into "how the predictors evaluated by the random forest work together" (Tagliamonte and Baayen 2012: 26).

Conditional inference trees, by contrast, constitute a very visual means of examining how predictors combine, yielding 'branches' that can be interpreted as '*if* ... *then* ... '-statements. This aspect of conditional inference trees is particularly important in our case, because our aim is not only to estimate the amount of variation explainable by the authors' idiolect, but also to test the hypothesis that this variation is due to different and personalised choices in the use of gerunds. Still, we have been very careful to consider the risks associated with the two types of models. We wish to clarify that we are well aware that conditional inference trees are not "representative trees" of a random forest (Gries 2019: 24–25), and we by no means consider the conditional inference trees generated to be a summary representation of a complementary random forest. Instead, we take them to be a supplementary method that can provide additional (but separate) insight into how the data can be partitioned (for each author) given a number of suggested predictors.

3.2 Language internal factors

3.2.1 Determiner

As explained in Section 2, the progressive variant ing-Ø did not emerge and become incrementally more frequent in all possible contexts simultaneously. Rather, it emerged in one type of context, and subsequently went through a lengthy diachronic process in which the conditions constraining its occurrence became more flexible. After closely examining diffusion of ing-Ø, Fanego (2004:50) concludes that its diffusion abided by two linguistic hierarchies.

The first hierarchy is the "hierarchy of relative nominality" (Fanego 2004: 38), which involves the type of determiner used in combination with the *ing*-nominal. Fanego explains that ing-Ø first emerged through reanalysis in determinerless contexts before spreading to contexts with a determiner. Fanego (2004: 38) then further subdivides contexts with a determiner into two types: "one where the determiner is a possessive, as in *his signing of the contract* (...), and one where it is an article, usually definite, as in *the signing of the contract*". Determiner use will be considered as an independent variable in the model to investigate whether and how determiner use helps condition the use of ing-OF and ing-Ø in our data set. The predictor levels, exemplified in Table 2, include the definite article (*the*), a possessive (e. g., *my, his, their*), and determinerless or 'bare' gerunds (the three main categories discussed by Fanego (2004), and four less frequent determiner types (i. e., demonstratives, indefinite articles, quantifiers such as *every*, and *no*).

Table 2: Levels of categorical predictor determiner (det).

det	ing-OF	ing-Ø
bare (ø)	Vices, like young Trees, the longer they are let grow, the greater difficulty there is in **ø** felling **of** them; (Roger Boyle, 1648)	… they enlarge the Orifice and increase the number, and take delight in **ø** torturing the poore Patient, whom they have in Cure. (Peter Heylyn, 1643)
poss	… which were represented and commemorated in **their** eating **of** that bread and drinking of that cup. (John Tillotson, 1683)	Besides, this glorious Majesty is himself present to behold his Worshippers in **their** worshiping him. (John Bunyan, 1679)
the	… not only in **the** stopping **of** the Play, but in **the** hanging up **of** the Poets. (John Dryden, 1683)	Nor are you happier in the relating or **the** moralizing your fable. (John Milton, 1660)
dem	But how will you juſtify **this** introducing **of** God's Name only (like a Fool in a Play) to make the Company laugh (Robert Boyle, 1648)	Now if **this** Hat – Honour and shewing the bare Head be an Invention of man (George Fox, 1677)
a	… it is rather a soiling then **a** fulfilling **of** mariage-rites (John Milton, 1643)	thou wert thereby kept from **a** further shedding the blood of thy soul (George Swinnock, 1659)
quant/no	primitive Church and Christians used to bow at **every** mentioninge **of** the name Iesus (William Prynne, 1636)	though Satan 's Messengers have told you, there is **no** hearing his Voice now-a-days (George Fox, 1673)

3.2.2 Function

Besides the hierarchy of relative nominality, Fanego (2004: 50) also uncovered a 'hierarchy of grammatical relations', explaining that ing-Ø "became available first as prepositional adverbials and oblique complements, then as core complements internal to VP, and last of all as subjects". In this study, we distinguished gerunds functioning as subjects, direct objects, subject complements, and prepositional complements (Table 3).

The 'grammatical relation' that contains by far the largest number of observations in our data set is that of prepositional complements. It is important to note here that prepositional complements can hardly be considered a homogeneous group, as different prepositions have started to 'allow' ing-Ø at different rates (De Smet 2008, De Smet 2013). Thus, rather than distinguishing only one level of 'prepositional uses', we distinguished six levels. The four most frequent prepositions – that is, *in, by, for* and *of* – are considered in four separate levels. The remaining prepositions were grouped according to whether they were the

Table 3: Levels of categorical predictor function (func).

func	ing-OF	ing-Ø
subject	[Launching **of** a Ship] plainly sets forth Our double State, by First and Second Birth. (John Flavell, 1664)	[Blessing and Praising God] is the most high and honourable act of worship (George Swinnock, 1672)
object	… they altogether forbid [preaching **of** the Doctrines of Grace], which Gods word commends unto us (William Prynne, 1640)	… the chief captain came, and they saw them, then they left [beating Paul], and here the men of Israel were cryed unto (George Fox, 1655)
subject complement (scomp)	Who as hee makes the first atchievement of Saint GEORGE, to bee [the killing **of** a burning Dragon] (Peter Heylyn, 1631)	That an Oath is nothing else, but [The asking or desiring a Divine Testimony] for the confirmation of th_ truth of our testimony (John Flavell, 1677)
frequent preposition (*in, by, for, of*)	First, the offense taken and given by the Papists amongst whom they live, **by** [their worshipping of images] (Thomas Fuller, 1639)	purge the distempered humours and save the much gangrend body, **by** [cutting some rotten and putrifide members off] (John Milton, 1641)
temporal preposition	**After** [the suspected Poysoning **of** his Father], not inquir'd into, but smother'd up (John Milton, 1649)	You see what success this Learned Critick has found in the World, **after** [his Blaspheming Shakespear]. (John Dryden, 1696)
other preposition	for that Scripture lay much upon me, **Without** [shedding **of** Blood] there is no Remission (John Bunyan, 1666)	Before it was so much as possible to have been Raised by the Dean, **without** [his knowing any thing of it], till so informed. (Thomas Pierce, 1683)

complement of a temporal preposition (e. g., *after, before, since*) or any other preposition (e. g., *from, through, without*).

3.2.3 Verb type

Finally, it has been argued that in Modern English, certain verbs types were more likely to occur as ing-Ø (Fonteyn 2019: 162). This is particularly the case for highly frequent verbs such as *have*, and light verbs (*do, take, make, give*, etc. (Huddleston and Pullum 2002: 290–296)). We distinguished three levels of verb types – possessive *have*, light verbs, and lexical verbs (Table 4).

Table 4: Levels of categorical predictor verb_type.

verb_type	ing-OF	ing-Ø
lexical	Do you not know that he may refuse to Elect who he will, without **abusing of** them? (John Bunyan, 1674)	… others make them groan, by **abusing** them to sin, and subjecting them to their lusts. (John Flavell, 1669)
light	Else how can that Assembly say AMEN at their Prayer or **giving of** thanks? (John Bunyan, 1683)	you are not in a Capacity then of Praying or **giving** Thanks, but of … Damning and Murdering (George Fox, 1674)
have	he had allmoſt nothing left remaining to him, but the empty title, the **having of** a Sword carried before him, and ſome other outward pomps of Court (Peter Heylyn, 1647)	he will take care for his Heritage, whether that murmuring complaining mind against Poor People of **having** so many Children (George Fox, 1683)

3.3 Language external factors

3.3.1 Individual

To account for the individuality of the authors, a categorical independent variable with 19 levels was added to the model. This allows the measurement of the variance explained by individuality as a whole, but also the extent to which this factor helps predict whether ing-OF or ing-Ø is chosen while other language-internal and external factors are controlled for. Furthermore, any interactions found in the conditional inference tree involving the predictor author may reveal idiosyncratic differences between (subsets of) authors.

3.3.2 Age and generation

In any study that wishes to scrutinize the language of individuals, it is crucial to also consider some higher-order sociolinguistic variables to determine whether particular grammatical choices are either truly individualized, or more accurately captured by means of larger (socio-linguistic) generalizations. At present, very little is known regarding any sociolinguistic variation in gerund usage in the seventeenth century, and hence it is difficult to determine which factors are worth considering in the multifactorial model. Furthermore, the overall social diversity of the authors included in this study is quite limited, which adds some constraints to the types of sociolinguistic predictors that can be considered.

However, there are two higher-order sociolinguistic factors that could potentially improve the predictive power of the multifactorial models and warrant further investigation.

First, it is interesting to investigate whether any generalizations can be made with respect to the generation of the author. As indicated in Section 3.1, the present data set includes the written output of author from three different generations (as defined in EMMA). As such, the multifactorial model includes a categorical predictor generation with three levels, which indicate whether the author has been born between 1599 and 1613 (G1), between 1621 and 1627 (G2), or between 1639 and 1644 (G3). The inclusion of generation as a predictor helps determine the predictive strength of an author's year (or decade) of birth in determining the grammatical choices the author makes with respect to the ongoing syntactic change. In parallel, to control for lifespan change, it is also worth considering the age of the author at the time the observations in the data set were uttered. It warrants a brief note that the predictor age added in the model set out in Section 4 is numeric rather than categorical. Still, because the models adopted in this study only work with categorical binary splits, age is discretized automatically by the models (e. g., utterance produced by authors under 36 vs. over 36 years of age; cf. Tagliamonte and Baayen 2012: 167).

3.3.3 Genre

Finally, the genre of the text in which the gerund appeared was also added as an independent variable. The importance of taking the context of production into account in sociolinguistic studies has been a concern of traditional variational sociolinguistics, where this type of variation was called *style variation* and was mainly measured as the attention paid to speech (Labov 1995) or as the type of interaction between speaker and hearer (Bell 1984). However, the importance of the context of situation in linguistic production, especially for written language, has not been explored in detail. Yet, extensive empirical evidence has been found, both synchronically (Biber 1988, Biber 1995, Biber 2012) and diachronically (Biber and Finegan 1989; Biber and Gray 2013), that *genre* (or *register*) variation is a very strong predictor of lexicogrammatical variation. Such evidence is also compatible with cognitive and exemplar-based theories of language, which predict that each linguistic construction will also contain associations with the context in which it has been encountered (Bybee 2010).

In EMMA, genre is modelled using three levels: text form, prototypical text category, and genre labels. The latter, the most detailed category, contains 18

general genre categories and each category has further differentiation in sub-genres (Petré et al. 2019). For our study, we focused on the two genres with the largest amount of available data: (argumentative or descriptive) prose and letters. These two genre categories are also useful because the communicative situation of a typical letter greatly differs from the communicative situation of a typical piece of prose (for example for being addressed to a person as opposed to an unenumerated recipient, and for being less likely to be carefully edited). However, note that the genre labels in EMMA were not assigned on the basis of an analysis of communicative situations and therefore we cannot exclude that some of the texts labelled as 'letters' were indeed produced as pieces of prose but observing the genre conventions of a letter.

4 Results

4.1 General random forest and conditional inference tree

The results of a random forest analysis are most commonly presented as a ranked list indicating the variable importance of the independent factors. The standard way of assessing variable importance of predictors in cforest follows the permutation principle of the 'mean decrease in accuracy'. However, the performance standard variable importance measure has been shown to be problematic when applied to models trained on unbalanced datasets where the class memberships of the dependent variable are not equally distributed (Janitza et al. 2013). Given that the class distribution of the dependent variables in the present data set is roughly 30–70 (OF-Ø), the measure implemented here is a more robust variable importance measure (Varimp_AUC{party}), for which the importance of each variable is calculated based on the Area Under the Curve (AUC). In a nutshell, the AUC represents an estimation of the probability that a randomly chosen observation from the response class ing-OF receives a higher probability for ing-OF than a randomly chosen observation from the response class ing-Ø. An AUC value of 1 indicates that the predictors in the model perfectly help classify the observations in the response class (OF or Ø), whereas a value of 0.5 indicates that in about half of the cases the model makes a wrong prediction for a randomly chosen observation. More formally, the variable importance (VI) of predictor j is computed as the average difference in AUC values when j is present or not (j) across all trees (N) the random forest comprises:

$$VI_j^{(AUC)} = \frac{1}{N}\sum_{t=1}^{N}(AUC_{tj} - AUC_{t\neg j})$$

The results of the AUC-based Variable Importance ranking are presented in Figure 3. What these figures reveal is that a language internal factor, determiner use (*det*), is by far the most important predictor when it comes to variation between ing-OF and ing-Ø in the seventeenth century. Other language internal factors, such as the function the gerund serves in the larger clause (*func*) and the type of verb with which the gerund is formed (*verb_type*), rank third and sixth respectively. Turning to the variable importance of the language-external factors, it is interesting to observe that the predictor *author* surpasses predictors such as *generation* and *age*, which are indeed tied to the individual, but aim to capture larger (sociolinguistic) generalizations. The finding that the predictor *genre* ranks last in terms of importance is remarkable, because grammatical variation tends to be substantially affected by variations in situational constraints. Yet, this result can be easily explained if indeed a considerable number of the letters in EMMA are very similar situationally speaking to prose, for example because they were written with the ultimate purpose of being published for an unenumerated audience. Overall, then, the variable importance of author shows that there is variability in the extent to which authors from the same generation or of the same age opt for ing-OF or ing-Ø in particular (grammatical) contexts.

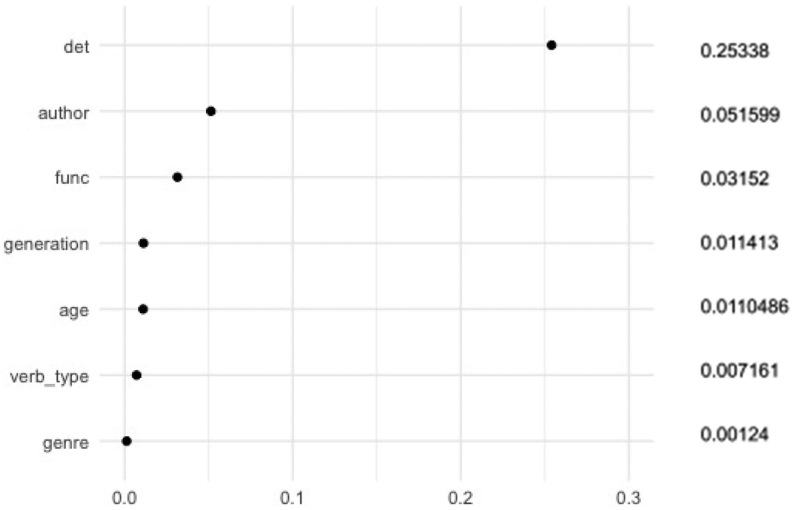

Figure 3: Variable importance ranking.

It should be noted here that the results of the random forest analysis give virtually no insight into which particular binary splits are made at what level in the 1,000 inference trees the random forest comprises. Furthermore, the results in Figure 3 are indeed indicative of individual variance, but they do not reveal anything about the direction of prediction or the interaction between predictors. To gain such insights, as explained in Section 3.1, a conditional inference tree can be grown. Generally speaking, random forests are better at capturing the relative explanatory power of each of the factors. This 'superiority' of random forest models is quite often illustrated by a large difference in classification accuracy between the single tree inference method and the random forest model (cf. Tagliamonte and Baayen 2012; Kerz and Wiechmann 2013; Szmrecsanyi et al. 2016). In the present case, however, the difference in performance between the single conditional inference tree and the random forest appears negligibly small (see Table 5). Both models were trained on the same random data sample (75%) and subsequently applied to a test set (25%) to calculate the models' Area under the Receiver Operating Characteristic (ROC) curve, accuracy (ratio of correctly predicted observations as ing-OF or ing-Ø to the total observations), and F1-score (the weighted average of the model's precision, i. e., the ratio of correctly predicted ing-Ø observations to the total predicted ing-Ø observations, and recall, i. e., the ratio of correctly predicted ing-Ø observations to all observations of ing-Ø):

Table 5: Model fits.

train (75%) + test (25%)	**cforest**	**ctree**
Area under the curve	0.8084	0.8037
Accuracy	0.8588068	0.8511364
F1-score	0.7346503	0.7247899

In total, the conditional inference tree algorithm produced a model with 39 recursive binary splits (9 levels, 79 nodes), in which all predictors occur at least once. A simplified version of the conditional inference tree (with the number of partitioning levels limited to 4) is presented in Figure 4. In line with the relative importance of determiner use yielded by the random forest algorithm, the data was first partitioned according to whether or not a particular set of determiners was used, with gerunds preceded by (in)definite articles, demonstratives, quantifiers, and *no* being separated from bare and possessive gerunds.

The splits at subsequent levels, then, represent further sub-partitionings of the data set after the previous split has been made. Consider, for instance, the

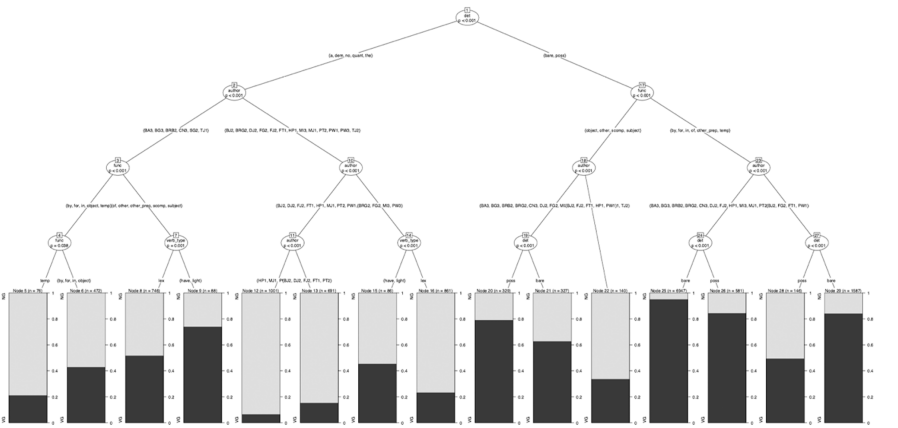

Figure 4: Conditional inference tree model with number of split levels set to 4.

subsection of the conditional inference tree singled out in Figure 5. The group bare and possessive gerunds (split made in Node 1) are further subdivided (Node 17) into those bare and possessive gerunds that function as the complement of a preposition and those that have more 'core' syntactic functions (i. e., subject, direct object, etc.). Within the group of bare and possessive gerunds in core functions (Node 18), then, the variation between ing-OF and ing-Ø is best explained by partitioning the data into tokens produced by John Bunyan, John Flavell, Thomas Fuller, Peter Heylyn and William Prynne versus all other authors, as the former group appears to be significantly more likely to use ing-OF under the conditions determined in earlier splits. The remaining set of authors (including all third-generation authors, but also second- and first-generation authors such as Robert Boyle and Jeremy Taylor) are significantly more likely to opt for ing-Ø in those contexts, particularly when the gerund is preceded by a possessive (Node 19).

4.2 Author-specific conditional inference trees

What is interesting about tree branches such as the one highlighted in Figure 5 is that they aid in highlighting how language-external factors and language-internal factors work together. As such, conditional inference trees allow us to examine in which respects all authors behave more or less homogeneously, but also where the grammatical choices of certain (subsets of) authors significantly deviate. To illustrate this, we can home in on some specific interaction effects to

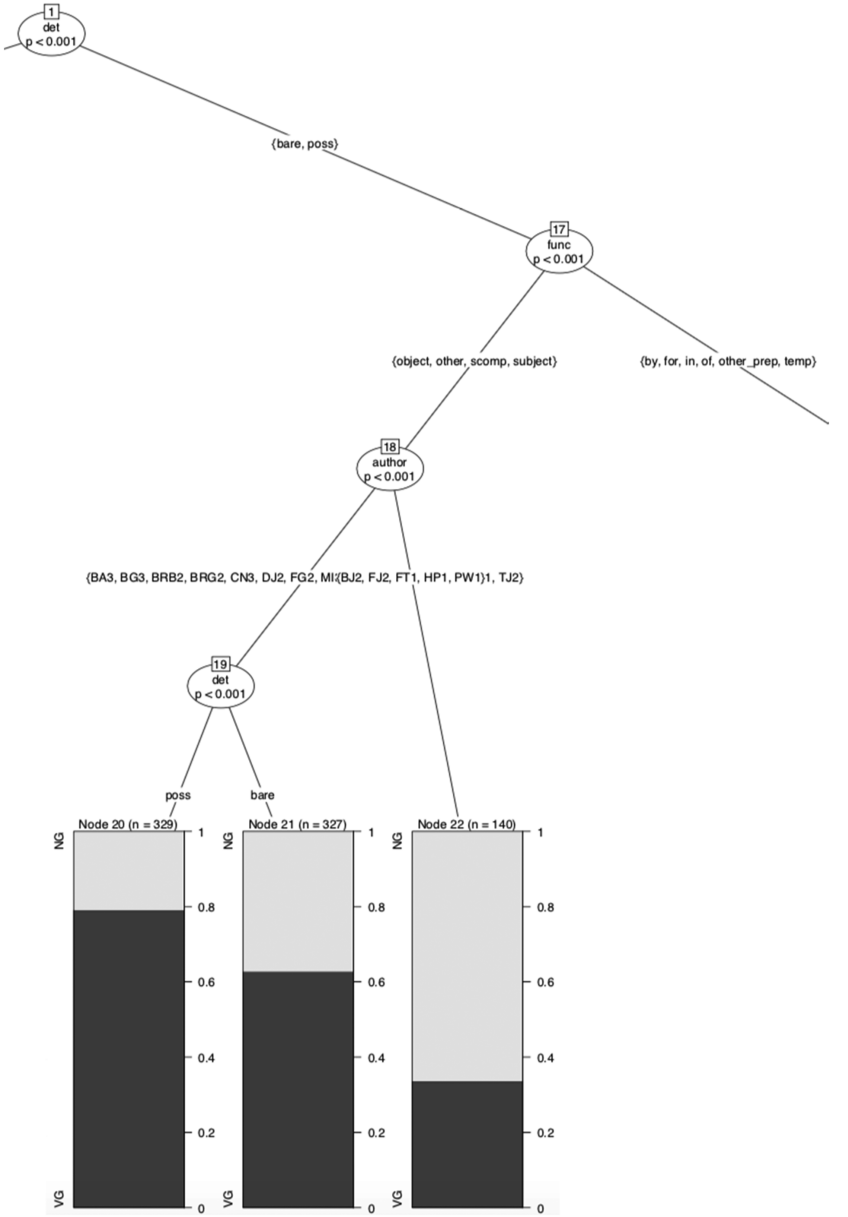

Figure 5: Branch of Figure 4 (including Node 1, 15, 16 and 17).

show how the idiolect of particular authors emerges as a function of the number of grammatical choices that they make. Following the interactions that shape the branch comprising of Node 1, 17, and 18, we can deduce that when a gerund without determiner which functions as the subject of the larger clause is used, John Bunyan is significantly more likely to use ing-OF than Robert Boyle. The difference between these two authors can be seen in the examples in Table 6, as well as in the two conditional inference trees in Figure 6, which have been grown following the same procedure described above but isolating only these two authors.

Table 6: Comparison of John Bunyan's and Robert Boyle's choice of gerund type in a comparable context (no determiner and functioning as subject).

Features:		**bare, subject John**	**(BJ2)** >	**2 tokens**	> **OF**
Authors:		**Bunyan Robert Boyle**	**(BRB2)** >	**2 tokens**	> **Ø**
BJ2	1685	Thus you see,		breaking	**of** bread, was the work
BJ2	1676	… when groundless considence,		pleading	**of** vertues … will not do
BRB2	1675	… in matters, where		embracing	or rejecting a course … is necessary
BRB2	1675	… the embracing … or not		embracing	this Religion, is an act of humane choice

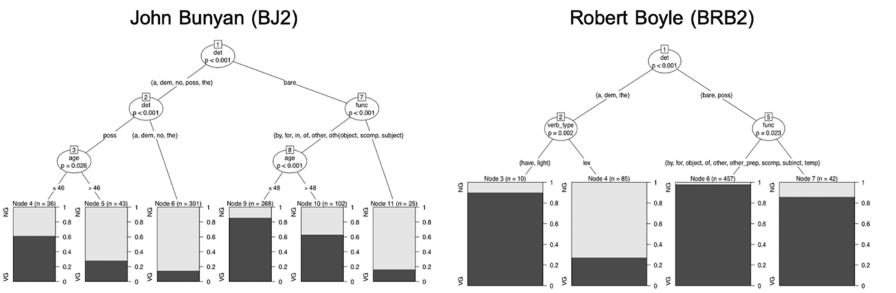

Figure 6: ctrees of John Bunyan and Robert Boyle.

These author-specific conditional inference trees are helpful in visualising the power of this method for research on individuality in language. For example, it is evident from Figure 6 above that, given the same context of a bare determiner in a subject position, Bunyan and Boyle behave very differently: the probability of encountering ing-OF in this context is much higher for Bunyan than for Boyle. Other things that

can be deduced are, for example, that *func* is a more important decision factor for Bunyan than for Boyle, and that *verb_type* significantly helps predict the variation for Boyle (when the gerund is preceded by *the, a, that* or *this*) but not for Bunyan. Moreover, because the variable *age* was also incorporated in the individualized trees, we can also highlight that John Bunyan took a retrograde turn to the conservative variant: in his later writings, Bunyan is more likely to use ing-OF following a possessive (Figure 6 – Node 5) and in bare prepositional complements, thus extending his general proneness to opt for ing-OF to a wider range of contexts.

Finally, it is also worth commenting on differences in the overall complexity or level of detail of the individualized trees of all authors. Some data sets, regardless of their size, simply require a smaller number of splits or a smaller range of predictors to effectively reduce variation. Figure 7 illustrates the individualized trees for two authors from the same generation, William Penn and Gilbert Burnett, who each contributed a large number of observations to the data set (2,377 and 2,457 tokens respectively).

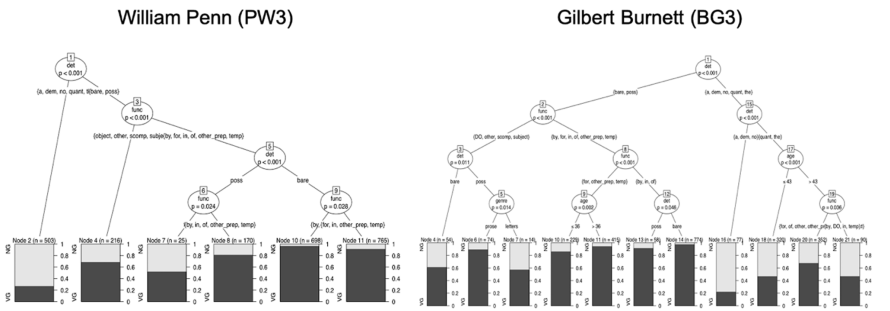

Figure 7: ctrees of William Penn and Gilbert Burnett.

The tree grown from William Penn's contributions comprises 11 nodes that solely involve binary splits according to determiner and function, whereas Gilbert Burnett's tree comprises 21 nodes involving splits by determiner, function and genre as well as some age effects (both retrograde and progressive).

4.3 Comparison of variable importance across author-specific models

As a final point of inquiry, a random forest (of 1,000 trees) was grown for each individual author. Whereas the general models summarised in Figures 3 and 4

present an aggregated picture of the factors associated with the use of ing-OF and ing-Ø for the whole set of authors, this analysis offers a comparison between each author's deducted partitioning models. The results of the individualized random forests are again presented as ranked lists of the independent variables included in the model according to their importance in predicting the variation between ing-OF and ing-Ø. However, to attain a detailed comparison between these author-specific models, the variable importance for the individualized random forests was calculated for each level of the categorical predictors. Note that the individualized random forests have AUC values that range from poor (between 0.65 and 0.70 for Aphra Behn, Gilbert Burnet and George Swinnock) to fair (between 0.77 and 0.79 for John Tillotson and Roger Boyle) to good (between 0.80 and 0.82 for the remaining 14 authors). Poor model fits are not due to issues of data quantity, but rather indicate that the proposed set of combined predictors is not optimally associated with the dependent variable for a small minority of authors.

Because importance scores cannot be compared across models in terms of their absolute value (Strobl et al. 2009: 336), Figure 8 presents each level (23 levels in total) in terms of its relative position in the variable importance ranking for each author (with 1 being the highest importance rank). With these individualized variable importance rankings, we can determine which particular predictor levels are (most) associated with the choice between ing-OF and ing-Ø for

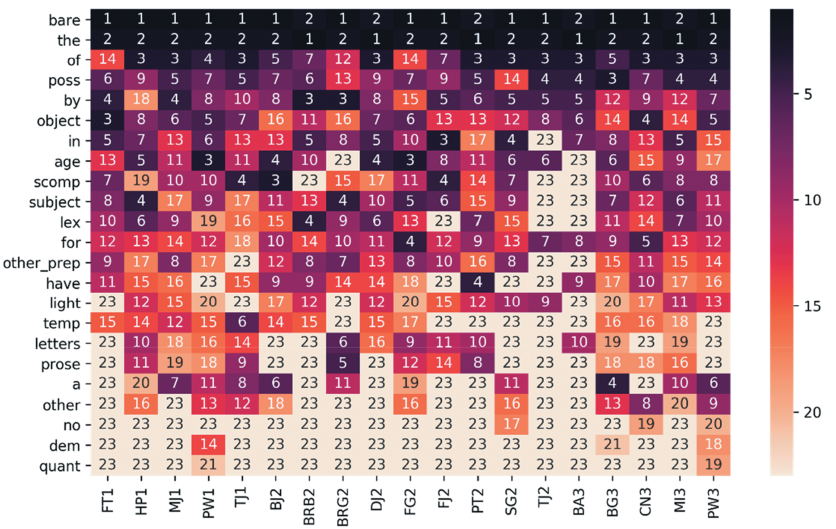

Figure 8: Author-specific variable importance rankings (AUC) per level.

each individual author. If the importance score indicated that the level did not contribute to predicting the observed variation (yielding a value below or equal to 0), its rank was set to 23 (last).

From these individualized rankings, a number of observations can be made. For all authors except Aphra Behn (BA3) and Roger Boyle (BRG2), there appears to be some predictive power associated with age, suggesting that it may be worth investigating lifespan change in gerund usage. The other levels in the rankings in Figure 8 can be interpreted differently, as they represent very specific grammatical contexts that are associated with either ing-OF or ing-Ø. What can be deduced from the rankings is that, for all authors, either *the* or *bare* is ranked first or second. As such, it appears that these two contexts constitute a shared grammatical constraint that is strongly associated with the two types of gerunds.

Yet, at the lower ranks, the author-specific models become more idiosyncratic: whether or not the gerund functions as a prepositional complement of *of* is ranked third in importance for only 58% of all authors, and similarities between authors drop to under 21% at all subsequent ranks. Furthermore, the individualized rankings again appear to differ in range and complexity (cf. Section 4.2): the model of John Tillotson (TJ2), for example, ranks only 8 levels (excluding age), whereas the model of his contemporary Roger Boyle (BRG2) ranks 16 levels. As a result, despite the fact that all models are similar at the highest ranks, none of the 19 author-specific models is identical in either the order or the range of ranked levels.

5 Discussion

The so-called 'gerund alternation', where speakers either opt for ing-OF (e. g., *kissing of frogs*) or ing-Ø (e. g., *kissing frogs*), is an interesting case to scrutinize, as earlier research already classified the diachronically unstable variation pair as one where the mixed linguistic output of individual language users generally centres around the population mean (Nevalainen et al. 2011; Baxter and Croft 2016). However, by examining the use of the two structural variants of the English gerund in the seventeenth century, this study demonstrated that there can still be heterogeneity within such a centralized population: individuals seem to have different systems to condition the variation they observe. Thus, our findings challenge the common belief that all constraints on grammatical variation are shared by all individuals in a community.

Further probing into the extent of individuality, it is interesting to observe that the general random forest (and conditional inference tree) presented in Section 4.1 indicated that the predictive power of individuality (author) surpasses that of other, more coarse-grained language-external factors tied to the individual, such as age or generation. Thus, the results seem to challenge the general practice of considering social and speech communities as a whole as the most important unit of analysis. Of course, some caution is necessary in treating the implications of these findings, as it is possible that relative importance of individuality can be explained by (one or multiple) latent social factors not included in the model. However, since we can infer that all authors are broadly comparable in terms of their level of education and social class based on their professions (Petré et al. 2019: 89), it is plausible to assume that the variance explained by the factor author is indeed capturing idiolectal differences.

In light of the model of language proposed by usage-based theories, these results can be explained by the fact that different individuals can come across different exemplars, and consequently will build slightly different *cognitive* (rather than or alongside social) models of a construction. Although conditional inference trees are a statistical and descriptive tool, the author-specific models as built in Section 4.2 – assuming that the most important predictors have been captured – can therefore also tentatively be interpreted as an approximation to their cognitive decision model, which should broadly reflect the exemplars that they have encountered. The findings presented in Section 4.3 also suggested that authors vary in terms of which contexts appear to matter most to them in deciding between ing-OF and ing-Ø. In sum, because these author-specific models are all different from each other in terms of order and breadth, it is not unreasonable to suggest that the authors under investigation appear to have had different understandings of the unstable gerund variable.

Yet, at the same time, there are clearly some generalizations that can be made across all authors in the sample. The results of Section 4.3 demonstrate that, for all authors, two contexts related to determiner use (*the* and *bare*) play the most important roles in the decision between the two types of gerund, which corroborates the results of the general models (Figure 3 and Figure 4) where *det* was found to be a stronger predictor than *author*. Overall, then, the picture emerging from these results is a complex one, which reveals that individuals appear to have idiosyncratic models of variation, but still behave homogeneously with regard to a *select* number of grammatical contexts. This complexity is due to the fact that the variety of analyses presented here do not only provide a measurement of the extent and importance of individual variance; they also allowed us to probe the locus of this individual variance. Whereas the most important language-internal conditions in which either ing-OF or ing-Ø is more

likely to appear are adopted by (and shared between) all authors, the observed inter-individual variation emerges at lower levels of importance, and in the range of ranked levels.

It is interesting to note here that these findings regarding the extent and locus of individuality are consistent with – but also help to explain – what we currently know about idiolect. Within forensic linguistics, casework and research have revealed how individuals can be distinguished from each other at fine levels of detail, such as the frequency of use of function words or combinations of words that might be very rare at the community level. One of the most significant problems in forensic linguistics – and in particular in computational authorship analysis – is that there is no full explanation for this finding. However, the methods and analyses presented here suggest a plausible hypothesis: the fluctuation in simple relative frequencies are largely predictable and ultimately depend on each individual's model of the variation, which is constituted by each individual's decision model. Thus, for gerunds, two authors' differences in the relative frequency with which they use ing-OF or ing-Ø can be explained by the different contexts in which they use them. Furthermore, our findings also suggest that, even if two authors select ing-OF or ing-Ø with (roughly) the same probability, they could have arrived at these distributions following different paths made by their personalised set of choices. These results are therefore compatible and lend evidence to Coulthard's (2004) hypothesis that an idiolect emerges from the co-selection of different options, which, in the present case, is constituted by the grammatical choices used as predictors in our models.

In sum, then, this study has offered further insight into "the extent and nature of individual variance for linguistic features at all levels of grammar" (Tagliamonte and Baayen 2012: 24) by providing further evidence that the generalizations, or rules of utterance production, "that individual members of the community have acquired" may differ between individuals within the same community (Dąbrowska Forthcoming). More specifically, what has been added to previous inquiries into the relation between individual language use and population-level language change is that, even within centralized or seemingly homogeneous populations, there can still be (subtle) heterogeneity in how individuals model grammatical variation with respect to specific (combinations of) conditions. Besides advancing our understanding of individuality in language, then, these results also offer a number of testable predictions and reveal some directions of further research. For example, the relative importance of age for different authors can be investigated further in a more developed sociolinguistic analysis, in order to offer an explanation for the presence or absence of (retrograde or progressive) age effects in specific conditions (cf. Anthonissen

Forthcoming). Furthermore, future research should investigate the extent to which the models of grammatical variation obtained by conditional inference trees are cognitively realistic using modern speakers and experimental settings, and conclusive evidence that the idiolectal effect in frequency of use is dependent on these cognitive models of grammatical variation can be tested by revisiting data from previous research in computational authorship analysis. At the same time, it may also be valuable to authorship attribution research to use the proposed tree-based methods with author as the dependent variable, to further pursue which (combination of) grammatical features perform best in distinguishing the writings of individual authors. Finally, in order to further improve explanatory accounts of idiolectal grammar, it is worth investigating *why* there is more or less inter-individual variation with regard to particular features and contexts/constraints in syntactic variation. Certainly, the degree to which individuals are unique in their own version of their grammar – and why individuals can be unique in that respect – is an understudied area of linguistics, and further steps need to be undertaken to arrive at a well-rounded answer to these questions.

References

Anthonissen, Lynn. Forthcoming. Cognition in construction grammar: Connecting individual and community grammars. *Cognitive Linguistics*. https://www.degruyter.com/view/j/cogl.ahead-of-print/cog-2019-0023/cog-2019-0023.xml.

Barlow, Michael. 2013. Individual differences and usage-based grammar'. *International Journal of Corpus Linguistics* 18(4). 443–478. doi:10.1075/ijcl.18.4.01bar.

Baxter, Gareth & William Croft. 2016. Modeling language change across the lifespan: Individual trajectories in community change. *Language Variation and Change* 28(2). 129–173. doi:10.1017/S0954394516000077.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language Learning* 59(1). 1–26.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13. 145–204.

Biber, D. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–37.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge; New York: Cambridge University Press.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65. 487–517.

Biber, Douglas & Bethany Gray. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2). 104–134.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45. 5–32.

Burrows, John. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3). 267–287. doi:10.1093/llc/17.3.267.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Claes, Jeroen. 2016. *Cognitive, social and individual constraints on linguistic variation*. Berlin: De Gruyter.

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

Dąbrowska, E. 2018. Experience, aptitude and individual differences in native language ultimate attainment'. *Cognition* 178(May). 222–235. doi:10.1016/j.cognition.2018.05.018.

Dąbrowska, Ewa. 2012. Different speakers, different grammars'. *Linguistic Approaches to Bilingualism* 2(3). 219–253. doi:10.1075/lab.2.3.01dab.

Dąbrowska, Ewa. Forthcoming. Language as a phenomenon of the third kind. *Cognitive Linguistics*.

De Smet, Hendrik. 2008. Functional motivations in the development of nominal and verbal gerunds in middle and early modern English. *English Language and Linguistics* 12(1). 55–102. doi:10.1017/S136067430700250X.

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

Eckert, Penelope. 2019. The individual in the semiotic landscape. *Glossa: A Journal of General Linguistics* 4(1). 14. 10.5334/gjgl.640.

Fanego, Teresa. 2004. On reanalysis and actualization in syntactic change: The rise and development of English verbal gerunds. *Diachronica* 21(1). 5–55. doi:10.1075/dia.21.1.03fan.

Feltgen, Q., B. Fagard & Jean-Pierre Nadal. 2017. Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science* 4(11). 170830.

Fonteyn, Lauren. 2017. The aggregate and the individual: Thoughts on how to make the best of 'bad data.'. *English Language and Linguistics* 21(2). 251–262.

Fonteyn, Lauren. 2019. *Categoriality in language change: The case of the English gerund*. Oxford: Oxford University Press.

Gries, Stefan Th. 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 0(0). doi:10.1515/cllt-2018-0078. http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2018-0078/cllt-2018-0078.xml (accessed 28 January 2020).

Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3). 251–270.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15. 651–674.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, Marianne, Sandra Mollin & Simone E. Pfenninger. 2017. *The changing English Language: Psycholinguistic perspectives*. Cambridge: CUP.

Janitza, Silke, Carolin Strobl & Anne-Laure Boulesteix. 2013. An AUC-based permutation variable importance measure for random forests'. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-119

Johnson, Alison & David Wright. 2014. Identifying idiolect in forensic authorship attribution: An n-gram textbite approach'. *Language and Law/Linguagem E Direito* 1(1). 37–69.

Johnstone, Barbara. 1996. *The linguistic individual. Self-Expression in language and linguistics*. Oxford, UK: Oxford University Press.

Kerz, Elma & Daniel Wiechmann. 2013. The positioning of concessive adverbial clauses in English: Assessing the importance of discourse-pragmatic and processing-based constraints'. *English Language and Linguistics* 17(1). 1–23.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1995. *Principles of linguistic change: Internal factors*. vol. 1. Oxford: Blackwell. https://www.wiley.com/en-us/Principles+of+Linguistic+Change%2C+Volume+1%3A+Internal+Factors-p–9780631179146.

Labov, William. 2012. What is to be learned: The community as the focus of social cognition. *Review of Cognitive Linguistics* 10(2). 265–293.

MacKenzie, Laurel. 2019. Perturbing the community grammar: Individual differences and community-level constraints on sociolinguistic variation. *Glossa: A Journal of General Linguistics* 4(1). 28. doi:10.5334/gjgl.622.

Maekelberghe, Charlotte. 2017. *The use of nominal and verbal gerunds in Present-day English. A multifunctional comparative analysis*. KU Leuven Phd Dissertation.

Meyerhoff, Miriam & James A. Walker. 2007. The persistence of variation in individual grammars: Copula absence in urban sojourners and their stay-at-home peers, Bequia (St Vincent and the Grenadines. *Journal of Sociolinguistics* 11(3). 346–366.

Mollin, Sandra. 2009. "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations'. *International Journal of Corpus Linguistics* 14(3). 367–392. 10.1075/ijcl.14.3.04mol.

Nevalainen, Terttu, Helena Ramoulin-Brunberg & Heikki Manilla. 2011. The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change* 23(1). 1–43.

Petré, Peter. 2017. The extravagant progressive: An experimental corpus study on the history of emphatic [BE V *ing*]. *English Language and Linguistics* 21(2). 227–250. doi:10.1017/S1360674317000107.

Petré, Peter, Lynn Anthonissen, Sara Budts, Enrique Manjavacas, Emma-Louise Silva, William Standing & Odile A.O. Strik. 2019. Early modern multiloquent authors (EMMA): Designing a large-scale corpus of individuals' languages. *ICAME Journal* 43(1). 83–122. doi:10.2478/icame-2019-0004.

Petré, Peter & Freek Van de Velde. 2018. The real-time dynamics of the individual and the community in grammaticalization. *Language* 94(4). 867–901.

Poplack, Shana & Sali Tagliamonte. 1991. African American English in the diaspora: Evidence from old-line Nova Scotians. *Language Variation and Change* 3(3). 301–339.

Robinson, Justyna A. 2011. A sociolinguistic perspective on semantic change. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 199–230. Berlin/Boston: de Gruyter Mouton.

Rutten, Gijsbert & Marijke van der Wal. 2014. Social and constructional diffusion: Relative clauses in seventeenth- and eighteenth-century Dutch'. In Ronny Boogaert, Timothy

Colleman & Gijsbert Rutten (eds.), *Extending the scope of construction grammar*, 181–206. Berlin: De Gruyter.

Schmid, Hans-Jörg & Annette Mantlik. 2015. Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles'. *Anglia* 133(4). 583–623. doi:10.1515/ang-2015-0056.

Smirnova, Elena. 2015. Constructionalization and constructional change: The role of context in the development of constructions. In Jóhanna Barðdal, Elena Smirnova, Lotte Sommerer & Spike Gildea (eds.), *Diachronic construction grammar*, 81–106. Amsterdam: John Benjamins.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods'. *Journal of the American Society for Information Science and Technology* 60(3). 538–556.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348. doi:10.1037/a0016973.

Szmrecsanyi, B., Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English'. *English World Wide* 37(2). 109–137.

Tagliamonte, Sali. 2013. Comparative sociolinguistics. In J. K. Chambers & Natalie Schilling (eds.), *The handbook of language variation and change*, 128–156. Malden, MA: Blackwell. doi:10.1002/9781118335598.ch6.

Tagliamonte, Sali & Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice'. *Language Variation and Change* 24(2). 135–178.

Traugott, Elisabeth C. & G. Trousdale. 2013. *Constructionalization and constructional changes*. Cambridge: Cambridge University Press.

Weinreich, Uriel, William Labov & Marvin I. Herzog. 1968. *Empirical foundations for a theory of language change*. Austin, TX: University of Texas Press.

Wright, David. 2017. Using word n-grams to identify authors and idiolects. A corpus approach to a forensic linguistic problem'. *International Journal of Corpus Linguistics* 22(2). 212–241. doi:10.1075/ijcl.22.2.03wri.