The aggregate and the individual: thoughts on what non-alternating authors reveal about linguistic alternations

1. INTRODUCTION

In one way or another, historical linguists have always been aware of the limitations inherent to working with linguistic data from bygone ages. One of the most substantial limitations, as Petré points out, is all speakers of a historical variant of a language are unavailable for psycholinguistic study, essentially leaving researchers with their written records as the sole data source. As such, historical linguists often find themselves taking the role of corpus linguist, trying to understand the workings of a language "by studying aggregate data that pools the productions of many speakers and writers – often across different media, genres, registers, and even across different time periods" (Arppe et al. 2010: 3). As Petré points out, the practice of studying language on this aggregate level has dominated the methodologies in historical linguistic studies, and very little attention is paid to the individual level.

This aggregate- level focus contrasts with more experimental approaches to linguistic variation and change, where the relation between individual participants and aggregate or population levels is of more central concern (e.g. Baayen et al. 2008; the studies reviewed in Scott- Philips & Kirby 2010). Drawing inspiration from experimental research, Petré develops a method of 'experimental corpus study', in which he uses a large set of "found" corpus data and uses a selected set so that it approximates "elicited" questionnaire data (Arppe et al. 2010: 7), aiming to make the most of the – sometimes frustratingly – 'bad data' we have (Labov 1994: 11). In this response article, I will suggest a complementary approach that can further aid to maximally exploit the found historical corpus data in front of us.

2. WHY INDIVIDUALS MATTER: MIMICKING ELLICITED HISTORICAL DATA

The method Petré presents is based on a previous study, Petré & Cuyckens (2008), in which a mirror set of two constructions was created based on aggregate data. In the present paper, however, Petré (this volume) take this method one step further and applies it to individual language users by compiling a corpus (of minimally 500,000 words) for each author. Each corpus was subsequently searched for progressive [BE V*ing*] followed by a random sample of 100 hits for each author. Subsequently, each randomly retrieved [BE V*ing*] is matched with a mirrored example that contains the same (or equivalent) verb lexeme in the simple present from the same corpus. The appeal of the individualized 'segregated' perspective is that it goes beyond descriptive observations about the language as an abstract object, and focuses on teasing out motivated decisions made by individual language users (cf. Hawkins 2004). The method is designed to select a set of utterances that are part of the entire body of written outputs of an individual, so that the resulting data set mimics a linguistic questionnaire that could reveal why individuals alternate between [BE V*ing*] and the simple present.

One point that I would like to stress in this response is that, while the method requires the analyst to selectively focus on those utterances where alternation is theoretically possible[1], it should not require them to be selective in terms of whether or not the individuals in fact

[1] Or, as Petré (this volume: page) puts it, the focus is on "contexts where the two are really competing which eachother".

alternate between the two forms. In the context of an actual linguistic experiment with a questionnaire, the possibility in principle always exists that a participant consistently opts for a single variant, which essentially also reveals something about how individual language users deal with alternations. If one is to approach an individual's corpus as a body of completed questionnaires, then, the analyst should allow that it is occasionally not possible to create a mirror set for a certain individual because the individual only uses, for instance, the simple present. Indeed, as Petré (this volume: page) remarks, the "main criterion for inclusion was the size of each author's corpus", and "[u]se of present tense progressive [BE Ving] was not a criterion". Thus, the fact that each corpus allows for sampling a mirror set is not due to the fact that the authors were a priori selected for having variable grammars, but because each author does in fact use [BE V*ing*].

As the methodological design does not include 'alternates between both variants' as a selection criterion, the analyst might very well be confronted with the fact that not all individuals alternate, which might raise a number of questions: what if it *is* the case that a small, considerable, or even large part of the population does not use one of the alternating forms? What kind of clues does that give the analyst about the organisational principles behind the alternation that they are studying? In what follows, I would like to dig into these questions by looking at a concrete example. With this example, I would like to show that it is in many ways undesirable to exclude non-alternating individuals from a study on cognitive motivations behind (changing) alternation pairs. If non-alternating authors are not taken into account, the analyst potentially runs the risk of disregarding an interesting chunk of information that can only be revealed by fully embracing the individual author level.

3. WHY ALL INDIVIDUALS MATTER: A CASE STUDY

The particular case I will be focusing on is the competition between two types of deverbal nominalizations in *-ing*: nominal gerunds (NG), illustrated in (1), and verbal gerunds (VG), illustrated in (2).

(1) a. (…) they shall be pleased to order the Witnesses to be collected, in *doing of which* there will be very little extraordinary Trouble or Expence. (PPCMBE, 1749)
b. What improvements might also be made are only here proposed to further trial, in order to *the having of roses*, and perhaps some other flowers. (PPCMBE, 1780)

(2) a. Ermine felt her imprudence in *having risked the betrayal* (PPCMBE, 1865)
b. I was in Court all that day, or the greater part, but I do not recollect *being examined* on Mr. Hooper's trial. (PPCMBE, 1817)

Interestingly, the two forms have presumably been competing over the same functional environments since approximately 1250, when the verbal gerund first arose (Jack 1988; Fanego 1996; Fanego 2004; De Smet 2008). The examples in (1)- (2) already indicate that the usage profile of nominal and verbal gerunds is not entirely the same: nominal gerunds cannot express secondary tense or voice distinctions and cannot be formed based on the stative verb *be*, while verbal gerunds seem to do so quite naturally (cf. (2a-b)). Yet, there is a considerable amount of functional overlap between the two forms. Consider for instance the

gerunds in (3), which are both formed with the base verb *eat* and both function as a prepositional complement of *in*.

(3) a. It was rarely observed of Philo, when Adam made that fond Excuse for his Folly in *eating the forbidden fruit* (PPCEME, 1630)
b. Sir, I never disobey my Father in any thing, but *eating of green* Gooseberries. (PPCEME, 1696)

In the literature on the role of competition in linguistic change, it has been suggested that functional overlap between two (or more) forms in the language system has either one of two outcomes (Traugott & Trousdale 2013: 18). The first outcome is replacement or substitution, with one (or several) of the competing constructions declining (Leech et al. 2009). The other outcome is that both forms are retained. The more traditional assumption when such retention is observed is that both forms continued to exist because the functional overlap between them was sorted out in order to maintain a one-form-one-meaning organisation in the language system (e.g. Mondorf 2011: 397, Nuyts & Byloo 2015: 34)[2]. In this scenario the competing constructions develop towards a division of labour, as each of the constructions involved comes to be preferred in particular functional niches (Torres Cacoullos & Walker 2009). As Petré's study indicates, the development towards such division of labour may proceed through various intermediate steps: the progressive first contrasted with the simple present because it was used as a (marked) extravagant form before it became fully associated with non-extravagant ongoing situations.

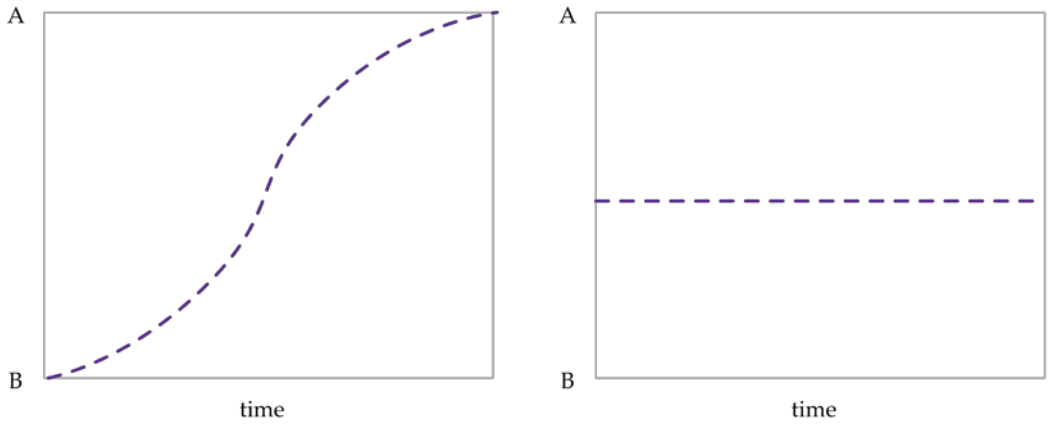

Figure 1 – Diachronic substitution and retention in a language.

Figure 1 schematically represents the different nature of substitution and retention processes. If a language contains a variant A and a variant B, which are in competition, the picture on the left illustrates a scenario where first only variant B exists, followed by an intermediate stage in which both variants occur. At the final stage, variant B has been replaced entirely by

[2] Note that more recently it has been argued that retention does not always mean that the constructions involved functionally differentiate over time: functionally overlapping syntactic variants can continue to stably co-exist for long stretches of time (Torres Cacoullos & Walker 2009; Noël 2003), or they can even become functionally more alike (De Smet *et al*. forthc.).

variant A. The picture on the right illustrates a hypothetical case of retention, in which variant A and B exist alongside each other with a relative frequency of roughly 50% for A and B. Both variants continue to co-exist as time progresses.

For nominal and verbal gerunds, De Smet (2008) has suggested that the rise of the verbal gerund constituted a large-scale replacement of nominal gerunds. The functional motivation behind this substitution is attributed to the fact that verbal gerunds are a short (more economic) means of expression, along with the fact that verbal gerunds are more syntactically flexible (De Smet 2008: 60, 95). A similar scenario is proposed in Nevalainen *et al.* (2011: 12), who suggest (albeit implicitly) that gerunds gradually transform from full abstract nouns with *of*-phrases into verbal structures. This process of substitution (or transformation) first affected nominal gerunds without initial determiners and in prepositional contexts (*by eating (of) the apple*, also see (3), (1a) and (2a)), and gradually spread to other contexts (Fanego 2004; De Smet 2013: 138). In more recent research, however, it has been argued that the verbalization of the gerund is perhaps not a clear-cut substitution process, but might in fact be a case of retention and niche-formation. Over the course of early and late Modern English, nominal gerunds have gradually specialized into their own functional niche (Fonteyn et al. 2015a, 2015b), resulting in a tendency towards diagrammatic iconicity with verbal gerunds replacing nominal gerunds in more clause-like functional domains, and nominal gerunds specializing to those functions that are more commonly associated with prototypical nouns (Fonteyn 2016).

Taken together, it appears that the formal and more recently also the functional-semantic history of the English gerund has been studied quite elaborately at the aggregate level. However, it is not necessarily the case that aggregate-level tendencies accurately (or even roughly) reflect the behaviour of individual language users. This is illustrated in the schematic representation of three hypothetical retention scenarios (through division of labour) in Figure 2. If we observe retention (through division of labour) on the aggregate level, it can of course reflect the behaviour of a homogenous group in which all individuals use the alternation (roughly assigning each form to its specific niche; illustrated in Figure 2, scenario 1).

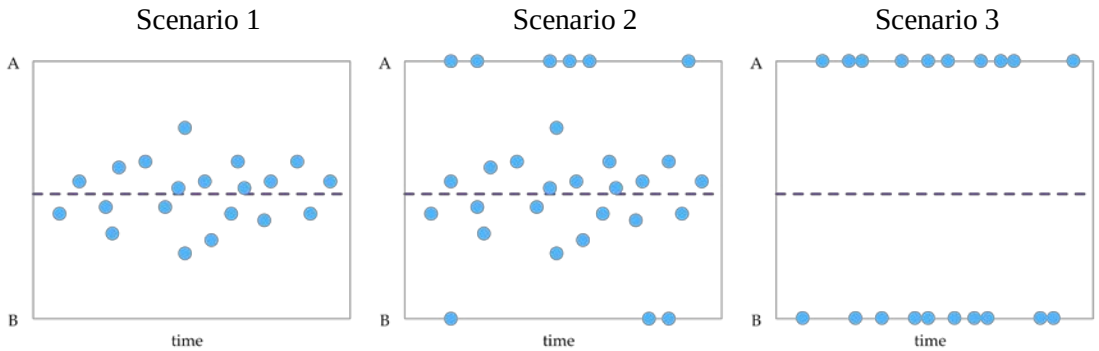

Figure 2 – Three artificial scenarios of alternation between variant A and B. The dashed line shows the aggregate level outcomes, which appear similar in each scenario. The blue dots represent the alternating behaviour of each individual that makes up the population.

In a less ideal case, however, the aggregate level division of labour only represents the behaviour of a part of the entire population. This is illustrated in Figure 2 (scenario 2), where a relatively small group of individuals appears to use only one variant, which can be either the

new form A (in their study of a range of diachronically changing alternations – including the gerund alternation – these are called “progressive individuals”, cf. Nevalainen *et al*. 2011) or the old variant B (“conservative individuals”, Nevalainen *et al*. 2011). Finally, in the most extreme case, all individuals are either progressive or conservative, as illustrated in Scenario 3. The problem in such a scenario is that the functional-semantic division of labour observed on the aggregate-level only exists in the language as an abstract object, and none of the individuals in the corpus actually uses both variants side by side. Consequently, one important question about the historical development of the English gerund that remains unexplored thus far is whether the emerging division of labour described by Fonteyn (2016) exists beyond the aggregate level, and describes an actual cognitive reality for individual language users. To address questions pertaining to the cognition of individual speakers and why the English gerundive system developed the way that it did in more depth, it is important that the aggregate level findings are considered in light of the behaviour and choices of individual language users.

2.1 DATA

In what follows, I will make a brief attempt to segregate the analysis of the aggregate- level findings of studies like that of Fonteyn (2016) by looking at the alternating behaviour of individual authors. Subsequently, I will raise a few questions about what the results mean for the study of alternations. The data presented here have been taken from the Penn Corpus of Early Modern English (PPCEME) and the Penn Corpus of Modern British English (PPCMBE). Both the PPCEME and the PPCMBE are divided into three 70- year subperiods. I will focus on four sequential time periods: 1570- 1639 (E2), 1640- 1709 (E3), 1710- 1779 (L1), and 1780- 1850 (L2). These subperiods are the same as those suggested by the corpus developers (Kroch *et al*. 2004; Kroch *et al*. 2010), and genre balance is roughly similar in each of them. For each periods I queried all words over four letters ending in *-ing* (string:\w{2,}ing\b) and subsequently filtered out all participial uses of V*ing* (progressives, premodifying present participles, etc.), all non- competing gerunds (formed with *be*, expressing passive voice or secondary tense, etc.), and other irrelevant hits in *-ing* (e.g. *according, painting, evening*, ...).

Table 1 lists the absolute frequency of nominal and verbal gerunds attested in the corpus, followed by the number of different authors in each period that used a gerund, the number of individuals that used both nominal and verbal gerunds, and the number of individuals that only used one of the two types. Note that the total number of authors in the Penn corpus for each subperiod is not the same as the total number of alternating and non-alternating authors considered in this case study. This has two reasons. First, a small set of authors did not use any type of gerund. In a questionnaire setting, these authors are the equivalent of participant that left the assignment blank. Because these authors did not yield any analysable data, they have been excluded from further consideration. Second, authors of whom the written output included less than 5 gerunds in total have been excluded to reduce the chance of spurious results.

Period	VG	NG	Authors total	Alternating	Non-alternating total	Only NG	Only VG
E2	596	611	48	36 (95%)	2 (**5%**)	1 (2.5%)	1 (2.5%)
E3	1263	309	59	38 (86%)	6 (**14%**)	0	6 (14%)
L1	1655	149	40	26 (65%)	14 (**35%**)	0	14 (35%)
L2	1253	140	28	24 (89%)	3 (**11%**)	0	3 (11%)

Table 1 - Frequencies per period

2.2 DISCUSSION

Figure 3 plots the alternating behaviour per individual over the periods under investigation. The smoothed line (Friedman's super smoother, cf. Friedman 1984) indicates the aggregate-level tendency. Importantly, Figure 3 only includes authors that have variable grammars and as such excludes all consistently progressive and conservative individuals.

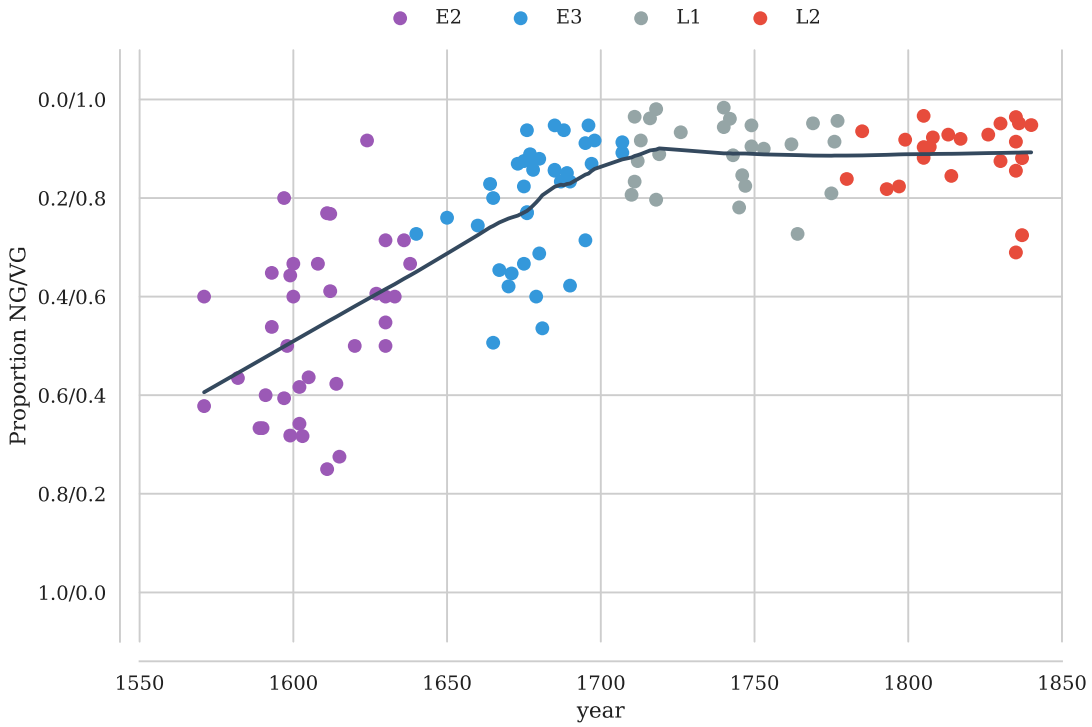

Figure 3 - Diachronic development of the nominal gerund/verbal gerund alternation. The dots represent individual authors. Dots towards the bottom of the graph represent authors who were more inclined to use nominal gerunds, whereas dots towards the top of the graph represent authors who are more inclined to use verbal gerunds. Non-alternating authors are not included.

In the early Modern Period (E2: 1570- 1639, E3: 1640- 1709), the development resembles a diachronic substitution process (with verbal gerunds replacing the nominal ones), but the proportion stabilizes at roughly 0.1/0.9 in the late Modern period. The vast majority of gerunds are verbal, but the nominal gerund seems to stand its ground in a small usage niche.

Figure 4, on the other hand, includes the entire set of authors. At first glance, the observed tendency seems largely similar to that in Figure 3. However, Figure 4 (as well as the frequencies in Table 1) indicates that, especially between 1640 and 1779, only a part of the entire population uses both nominal and verbal gerund (cf. Scenario 2 in Figure 2). While the number of authors solely using nominal gerunds is quite small and limited to the first period (one author providing more than 5 tokens, 2.5% of the population between 1570-1639), as much as 35% of the entire set of authors uses only verbal gerunds between 1710 and 1779[3]. Put differently, if we attest any sort of functional division of labour between nominal and verbal gerunds in this period, we must be aware that only 65% of the individuals uses both forms.

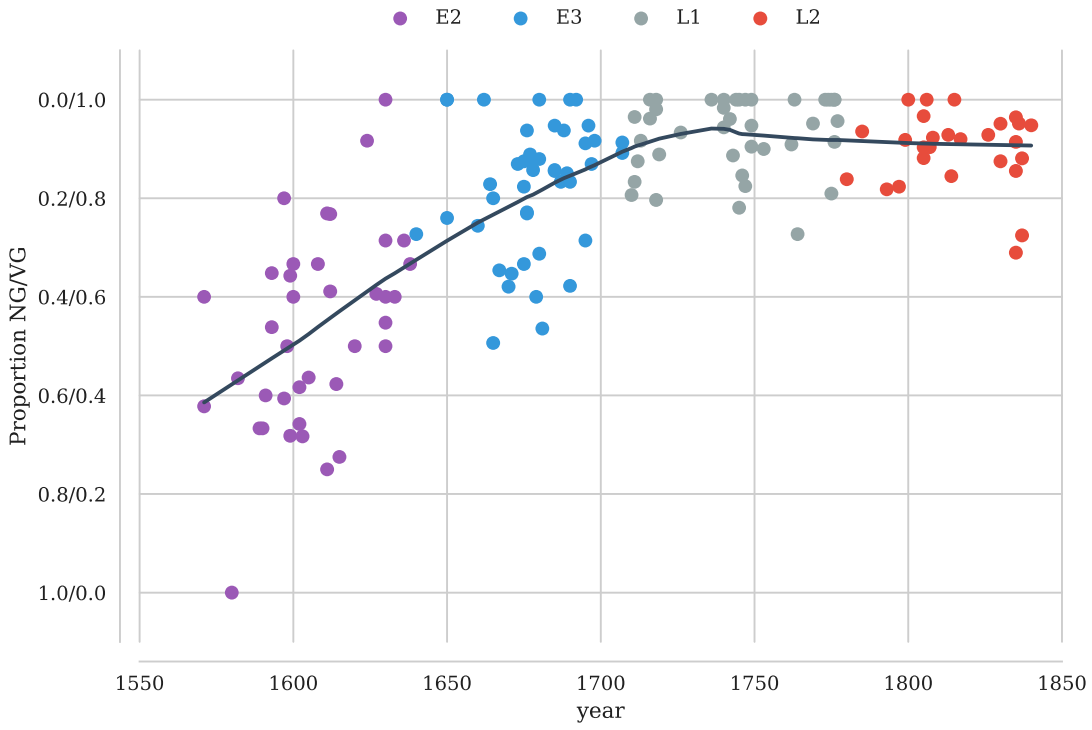

Figure 4 – Diachronic development of the nominal gerund/verbal gerund alternation (cf. Figure 3). Non-alternating authors have been included.

When such figures are revealed, the question arises whether one throws out the baby with the bathwater if these non- alternating authors are excluded from the data. The answer to this question is that it depends on a few factors. A closer inspection of the set of non- alternating individuals might reveal that the apparent non-alternating authors are a consequence of 'bad

[3] When the minimum token threshold is raised to a minimum of 10 tokens per author, the percentage of non-alternating authors remains 35% of the population. When the threshold is raised to 25 tokens, the proportion of non-alternating authors is still 30% of the population. The full data set, including the number of tokens for each author, is available on https://github.com/LFonteyn/ELLFonteyn2017.

data' because the absolute frequency of tokens that they have produced is far from substantial. One way to control for this is to set a minimum number of tokens threshold for each author, or to make sure that a sufficiently large corpus is compiled for each author (cf. Petré). Another possible factor to take into account might be that all non- alternating authors have produced texts from a similar genre, which might reveal a genre effect the analyst has to take into account. But if no such circumstances exist, the group of non- alternators might launch more substantial issues. If we find that a (large) share of the population does not abide by the rules you have set out, the motivations and cognitive principles you have uncovered might well be not universal. What does that say about the cognitive reality of the underlying motivations you have discovered? And if we find that the principles that we have described are not universal, how much does that matter?

In the case of the gerund- alternation, I believe that it is perfectly possible that the fraction of non- alternating authors (listed in Table 1 and Figure 5) reveals something about the status of the division of labour described in for instance Fonteyn (2016) in each of the individual time periods. In choosing between linguistic variants, language users are often driven by competing motivations (MacWhinney *et al.* 2014): on the one hand, principles of economy motivate speakers to opt for the shortest means of expression, while on the other hand iconic principles drive language users to opt for the form that most accurately resembles other linguistic forms in its paradigm (i.e. 'system pressure', Haspelmath 2014). As already indicated by De Smet (2008), the rise of the verbal gerund initially constituted a substitution process driven by forces of economy. We could speculate that the growing rate of progressive authors solely using verbal gerunds between 1570 and 1779 might be a reflection of the initial victory of economic motivations.

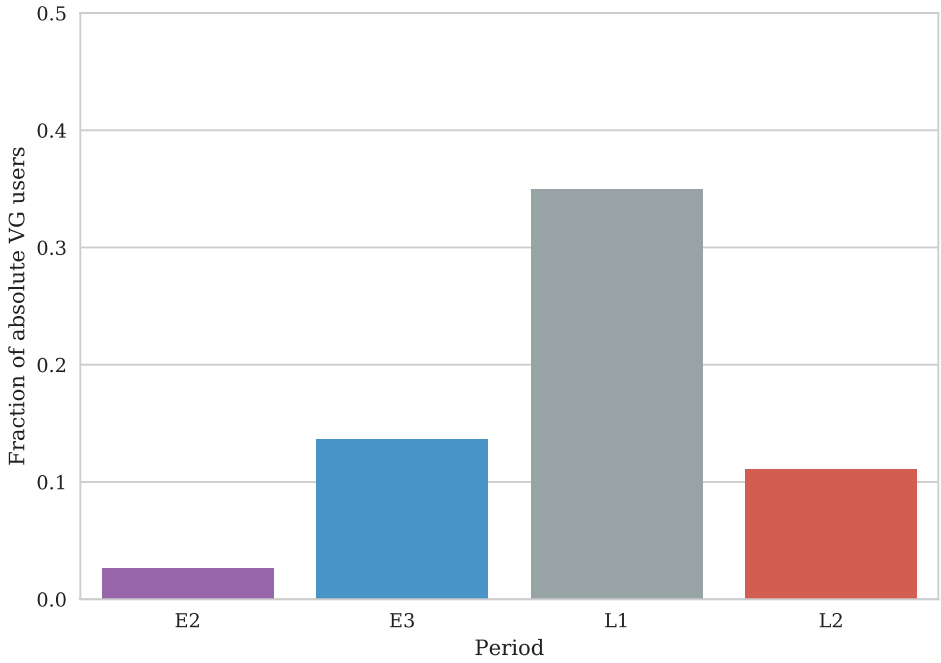

Figure 5 - Diachronic changes in the fraction of non-alternating author solely using verbal gerunds

In the final late Modern period (1780-1850), then, the fraction of non-alternating authors again decreases. While more detailed research on a larger data set is desirable, this could tentatively indicate that by the end of the 18th century, other motivations have counteracted the economy-driven substitution process, the majority of language users now starting to assign nominal and verbal gerunds to their own (iconically motivated) niche.

3. CONCLUSION

As aptly indicated by Petré, aggregate-level studies treating language as an abstract object do not fully comply with the idea that linguistic change is driven by the repeated behaviour of individual language users (see also Croft 2000; Scott-Phillips & Kirby 2010). A potentially problematic consequence of studying alternations only on the aggregate level is that one might uncover cognitive-functional divisions of labour between the alternating forms that do not actually characterise motivations adopted by individuals (as in scenario 3, in figure 1, repeated in Figure 6 below). As the individual mirror set method presented by Petré focuses on individuals that use both forms simultaneously, it prevents researchers from drawing such problematic conclusions. As it turns out, the linguistic change described by Petré seems to fit the profile of Scenario 1, in which all authors alternate in some way or another between progressive [BE V*ing*] and the simple present. The case presented in this response paper, however, seems to be more in line with Scenario 2, in which the population comprises alternating individuals as well as progressive (and conservative) individuals that consistently use only one variant.

As I see it, the case study and reflections presented in this response suggest two action points for the research agenda of the cognitive historical linguist. First, following Petré, I believe that we have to embrace the fact that there are plenty of *I*'s in *team*. Historical linguists interested in revealing cognitive motivations behind linguistic alternations should more often adopt an approach that compares and contrasts aggregate-level with individual-level findings. Second, when studying the diachronic development of alternations on the individual level, we should be careful not to filter out what might be useful data. Reporting the fraction of non-alternating individuals and including their linguistic outputs into the data set is a useful practice, which either supports or raises doubts about the universality of the cognitive principles the study unveils. Either way, it brings us closer to the truth.

REFERENCES

Arppe, Antti, Gilquin, Gaëtanelle, Glynn, Dylan, Hilpert, Martin & Arne Zeschel. 2010. Cognitive Corpus Linguistics: five points of debate on current theory and methodology. Corpora 5.1, 1-27.

Baayen, Harald, Davidson, Donald J. & Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390-412.

Croft, William. 2000. Explaining language change: An evolutionary approach. London: Longman.

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.

De Smet, Hendrik. 2013. Spreading Patterns: Diffusional Change in the ENglish System of Complementation. Oxford: Oxford University Press.

De Smet, Hendrik, D'Hoedt, Frauke, Fonteyn, Lauren & Kristel Van Goethem. Forthc. The changing functions of competing forms: attraction and differentiation. Cognitive Linguistics.

Fanego, Teresa. 1996. The gerund in Early Modern English: Evidence from the Helsinki Corpus. Folia Linguistica Historica 17, 97- 152.

Fanego, Teresa. 2004. On reanalysis and actualization in syntactic change: the rise and development of English verbal gerund'. Diachronica 21.1, 5- 55.

Fonteyn, Lauren, De Smet, Hendrik & Liesbet Heyvaert. 2015a. What it means to verbalize: the changing discourse functions of the English gerund. Journal of English Linguistics 43.1, 1- 25.

Fonteyn, Lauren, Heyvaert, Liesbet & Charlotte Maekelberghe. 2015b. How do gerunds conceptualize events? A diachronic study. Cognitive Linguistics 26.4, 583- 612.

Fonteyn, Lauren. 2016. Categoriality in language change: the case of the English gerund. PhD dissertation, department of linguistics, KU Leuven.

Friedman, Jerome H. 1984. A variable span smoother. Technical Report 5. Laboratory for Computational Statistics, Department of Statistics, Stanford University.

Haspelmath, Martin. 2014. On system pressure competing with economic motivation. In MacWhinney *et al*. Competing motivations in grammar and usage,197- 208. Oxford: Oxford University Press.

Hawkins, John A. 2004. Efficiency and complexity in grammars. Oxford: Oxford University Press.

Jack, George. 1988. The origins of the English gerund. Nowele 12, 15- 75.

Kroch, Anthony, Beatrice S., & Lauren Delfs. 2004. Penn-Helsinki parsed corpus of Early Modern English. www.ling.upenn.edu/hist- corpora/PPCEME- RELEASE- 1

Kroch, Anthony, Santorini, Beatrice S., & Ariel Diertani. (2010). Penn Parsed Corpus of Modern British English. http://www.ling.upenn.edu/hist- corpora/PPCMBE- RELEASE- 1

Labov, William. 1994. Principles of Linguistic Change. Volume 1: Internal Factors. Oxford: Blackwell.

Leech, Geoffrey, Hundt, Marianne, Mair, Christian & Nicholas Smith. 2009. Change in Contemporary English : a grammatical study. Cambridge: Cambridge University Press.

MacWhinney, Brian, Malchukov, Andrej & Edith Moravcsik (eds.). 2014. Competing motivations in grammar and usage. Oxford: Oxford University Press.

Mondorf, Britta. 2010. Variation and change in English resultative constructions.

Language Variation and Change 22.3, 397- 421.

Noël, Dirk. 2003. Is there semantics in all syntax? The case of accusative and infinitive constructions vs. that- clauses. In Günther Rohdenburg and Britta Mondorf (eds). Determinants of grammatical variation in English, 329- 345. Berlin: Mouton de Gruyter.

Nuyts, Jan & Pieter Byloo. 2015. Competing modals: Beyond (inter)subjectification. Diachronica 32.1, 34- 68.

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

Petré, Peter & Hubert Cuyckens. 2008. Bedusted, yet not beheaded: The role of be-'s constructional properties in its conservation. In Bergs & Diewald (eds.), Constructions and Language Change, 133-170. Berlin: Mouton de Gruyter.

Scott-Phillips, Thomas C. & Simon Kirby. 2010. Language evolution in the laboratory. *Trends in Cognitive Sciences* 14, 411-417.

Traugott, Elisabeth Closs & Graeme Trousdale. 2013. Constructionalization and Constructional Changes. Oxford: Oxford University Press.

Torres Cacoullos, Rena & James A. Walker. 2009. The present of the English future: grammatical variation and collocations in discourse. Language 85, 321-354.