Benedikt Szmrecsanyi*

About text frequencies in historical linguistics: Disentangling environmental and grammatical change

DOI 10.1515/cllt-2015-0068

Abstract: This paper is concerned with the limitations of inferring grammar change from variable text frequencies in historical corpus data. We argue that fluctuating frequencies of grammatical variants in real time are a function not only of changing grammars but are also conditioned by what we call 'environmental' changes (for example, content changes) that affect the textual habitat. As a case study, we explore the English genitive alternation in the Late Modern English period and demonstrate that the English *s*-genitive is and always has been preferably used with animate possessors; if for some reason animate NPs are rare in some specific historical period or text, this will trivially depress *s*-genitive rates and boost *of*-genitive rates. Against this backdrop, the paper advocates probing the probabilistic underpinning of grammatical variability in diachrony, for the sake of keeping apart trivial habitat-induced frequency change and grammar change proper.

Keywords: frequency, probabilistic, genitive, variation, Late Modern English

1 Introduction

This study addresses the problematic status of text frequencies as a diagnostic marker of grammar change in corpus-based, variationist research designs. Smoke detectors go off when they detect smoke. But when a smoke detector goes off, this does not mean that the smoke detector has changed; what has changed is the smoke detector's environment, or habitat. We submit that corpus-based historical linguists often face a similar issue, in that fluctuating frequencies of grammatical variants are a function not only of changing grammars but are also conditioned by environmental changes in the textual habitat. So the crucial problem is that diachronically variable text frequencies often entangle environmental differences and grammatical changes. To disentangle the two types of change, we will argue that instead of focusing solely on text frequencies

***Corresponding author: Benedikt Szmrecsanyi,** Department of Linguistics, KU Leuven, Blijde-Inkomststraat 21, PO Box 03308, B-3000 Leuven, Belgium, E-mail: benszm@kuleuven.be

(How *often* do language users use some linguistic variant?), analysts need to explore the possibly historically evolving probabilistic conditioning of variants (*Why* do language users use the variants that they use?). This approach yields a more reliable diagnostic of grammar change.

Our proposal is not actually a very original one – variationist sociolinguists, for example, have been long aware that language change may manifest itself in extremely subtle shifts in the stochastic effects of conditioning factors, and that mere text frequencies (or variant rates, for that matter) may be as much about culture as they are about language. Yet in the corpus-based historical linguistics community, we seem to be dealing with a deeply entrenched reliance on the diagnostic power of corpus frequencies. We hasten to add that all other things being equal, diachronically fluctuating text frequencies of grammatical forms may indeed point to grammar change – but especially in historical linguistics, all other things are rarely equal. This is the central point that this paper seeks to emphasize.

This paper is structured as follows: In Section 2, we conduct a thought experiment to illustrate the problem. Section 3 further sets the stage by offering some crucial definitions and by presenting a very concise review of the relevant literature. Section 4 discusses variability between the *s*-genitive and the *of*-genitive in the Late Modern English period as a case study that highlights the limited potential of text frequencies as a diagnostic of grammar change. Section 5 offers some concluding remarks.

2 A thought experiment

Let us assume an ancient culture with some language *X* that has (at least in principle) an overt future marker *F*. The historical record that survives consists exclusively of one particular text type *R*. Assume further that before time t_1, a religious norm outlaws talk about the future in this particular text type. At time t_1, though, this norm is rescinded. The result is a beautiful *s*-curve-shaped frequency explosion of future marker *F*. Consider Figure 1, which plots hypothetical text frequencies of *F* on the *y*-axis[1] against hypothetical units of time on the

1 Note that for our thought experiment and for the argument in this paper, it does not matter at all if we consider absolute frequencies (e.g. frequency per million words of running text) or some relative frequency measure (e.g. the rate of *F* usage vis-à-vis usage of some other grammatical marker, such as the present tense).

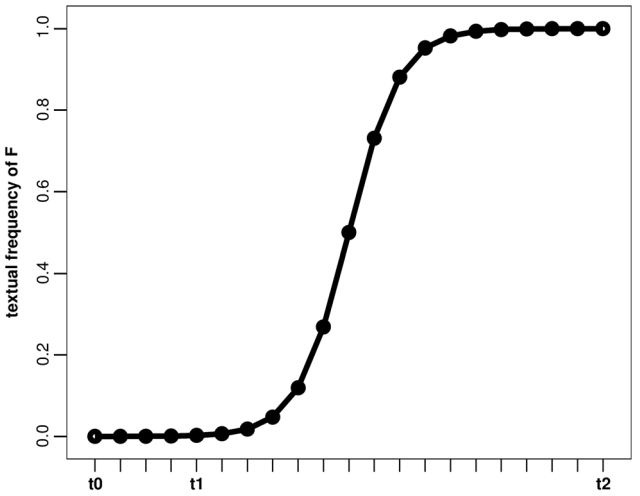

Figure 1: Hypothetical frequencies of future marker *F* (*y*-axis) at times t_0, t_1, and t_2 (*x*-axis).

x axis. Prima facie, the curve in Figure 1 looks like a language change phenomenon of the sort that every historical linguist is dreaming of discovering once in his/her lifetime. But we know how this curve came about, and so we must wonder: is the frequency change depicted in Figure 1 a symptom of grammar change, or should we be careful and avoid the term "grammar change" in this context?

This question is a rhetorical one, of course. Few analysts, whether of the formalist or functionalist persuasion, would regard the curve in Figure 1 as having anything to do with grammar change – rather, it depicts the time course of an environmental change that has altered the subject matters in one particular text type. It is clear that in the long run the frequency change depicted in Figure 1 may very well have linguistic and/or grammatical consequences (along the lines of Bybee 2006), but in the short run the curve in Figure 1 is what it is: trivial in linguistic terms.

3 Preliminaries: Definitions and literature review

The present study falls within the remit of variationist linguistics in a modern and fairly broad sense (Labov 1982; Tagliamonte 2001; Bresnan and Ford 2010). This means that we follow sociolinguistic theory and recent probabilistic approaches to language in assuming that grammatical variation and change

are often, and maybe even typically, gradient rather than categorical in nature. But what is "grammar"? We define grammar as the knowledge of

(i) *structural units*, however one's theoretical framework may conceive of them – symbols (e.g. verbs), constructions (e.g. the *s*-genitive construction), or (socio)linguistic variants and variables[2] (e.g. the active-passive alternation);
(ii) *probabilistic constraints*, which may be unit-specific (e.g. kinship relations favoring the English *s*-genitive) or general (e.g. long constituents tending to follow short constituents, at least in languages like English or German).

Note that probabilistic constraints may come with $p = 1$ (i.e. certainty), in which case we are not dealing with choice processes but with categorical constraints fully compatible with categorical notions of grammar in the spirit of, for instance, Chomsky (1965) (for example, *in subordinate clauses, speakers of German use verb-final word order with probability 1*).

So, against this backdrop, what is grammar *change*? Trivially, a grammar change may change the set of basic structural units; so, for example, lexical units may develop into new grammatical units. This sort of change is extremely well studied (e.g. Hopper and Traugott 2003). But we will assume that grammar change may also be probabilistic in nature, altering the *constraint set-up* – and possibly *ranking* – that fuels linguistic choice making. Crucially, probabilistic change ought to be *habitat-independent*, in that it should be a general change observable across text types. This is another way of saying that the frequency profile of grammatical variants generated by the alleged change should not be tied to, or predictable on the basis of, particularities of a specific text or genre. We further stipulate that genuine grammatical change should operate *below the level of conscious awareness* (parlance of Labov 1972), at least in its initial stages. Sapir's notion of "drift" seems apropos:

> The drift of a language is constituted by the *unconscious* selection on the part of its speakers of [...] variations [...] In the long run any new feature of the drift becomes part and parcel of the *common, accepted speech.* (Sapir 2004: 127; emphases mine)

Given these criteria, we stress that the frequency change in our thought experiment (Figure 1) clearly does not qualify as grammar change. First, the frequency

2 Contra Lavandera (1978), we follow, e.g., Weiner and Labov (1983) in assuming that the notion of the linguistic variable can be extended beyond the phonological level.

explosion in our thought experiment is tied to one particular text type in which for non-linguistic reasons, usage of the future marker was non-existent before time t_1. Second, the booming usage of the future marker after time t_1 is clearly not engendered by "drifty" linguistic choices below the level of consciousness. Instead, our hypothetical writers consciously chose to write about new subject matters (that is, the future); this was an environmental change that secondarily triggered increased text frequencies of future marker *F*.

To what extent is the foregoing discussion relevant to current studies in corpus-based historical linguistics? We begin by acknowledging the primacy of speech: the most appropriate text type to study grammar change is, or would be, everyday face-to-face interaction, a genre that drives change but is simultaneously fairly immune to stylistic fads and fashions (for example, Paul 1920: 32). In our thought experiment, we may suppose that speakers of language *X* presumably had always used future marker *F* in ordinary conversation (remember that it was text type *R* only which was subject to religious censorship). The problem, of course, is that the long-term historical record that we have in historical linguistics at the present time does not document speech. In the absence of historical records of face-to-face interaction, however, the dynamics of written registers (read: environmental changes affecting the textual habitat) are a serious confounding variable – much as in our thought experiment, the norm outlawing talk about the future is a confounding factor. Again, we stress that this problem is in principle well-known (see, e.g., Biber and Finegan 1989; Hundt and Mair 1999; Biber and Conrad 2001). Yet, there has been a tendency to shrug off the problem, and/or to concede defeat by accepting that trying to disentangle grammar change from environmental change (such as dynamically evolving written text types) is a hopeless endeavor (consider, for example, Curzan 2009: 1103). The present study is an extended empirical argument that we can do better than that, and that it is in fact possible to disentangle environmental change (for example, fads and fashions affecting a written register) from grammar change proper. To do so, though, we need to go beyond mere discussions of text frequencies.

4 A case study: Genitive variability in Late Modern English

To furnish a little case study, we discuss in what follows variability between the *s*-genitive (as in (1a)) and the *of*-genitive (as in (1b)) in the Late Modern English period:

(1) a. *before [**the Seneschal**]s [**Brother**] could arrive, he was secured by the Governor of Newport* <1682pro1.n2b>[3]
b. *the Duke of Norfolk, having lately received another Challenge from [**the Brother**] of [**the Seneschal**], went to the place appointed* <1682pro1.n2b>

The genitive alternation is a syntactic alternation that affects the order of the so-called possessor (*the Seneschal*) and the so-called possessum (*(the) brother*).

4.1 The history of genitive variability

Historically, the *of*-genitive is the incoming form, which appeared during the ninth century. According to Thomas (1931: 284) (cited in Mustanoja 1960: 75), the inflected genitive vastly outnumbered the periphrasis with *of* up until the twelfth century. In the Middle English period, we begin to observe "a strong tendency to replace the inflectional genitive by periphrastic constructions, above all by periphrasis with the preposition *of*" (Mustanoja 1960: 70), to the extent that the inflected genitive came close to extinction (Jucker 1993: 121). The frequencies calculated by Thomas (1931) show that by the fourteenth century, the *of*-genitive had a market share of about 84%. Yet in the Early Modern English period, we see a revival of the *s*-genitive, "against all odds" (Rosenbach 2002: 184). In Present-day English, empirical research has reported comparatively high frequencies of the *s*-genitive (for example, Rosenbach 2002; Szmrecsanyi and Hinrichs 2008). The consensus is that the *s*-genitive is spreading right now (for example, Potter 1969; Dahl 1971; Raab-Fischer 1995; Rosenbach 2003; Szmrecsanyi 2009, 2010).

4.2 Data

We re-analyze the genitive dataset presented in Wolk et al. (2013) (which in turn partially overlaps with the dataset investigated in Szmrecsanyi 2013). The dataset is drawn from ARCHER, *A Representative Corpus of Historical English Registers*, release 3.1 (Yánez-Bouza 2011). ARCHER covers the period between 1650 and 1999, spans about 1.8 million words of running text, and samples a number of different registers. We shall restrict attention to *s*-genitives and *of*-genitives in ARCHER's British English news (a fairly "agile" genre (Hundt and Mair 1999: 236)) and letters section (a register that is considered fairly oral, at least in regard to

3 All linguistic examples in this paper are drawn from the ARCHER corpus (see Section 4.2) and are referenced by ARCHER text identifiers.

private letters (Raumolin-Brunberg 2005: 57)). Our textual database thus comprises 257 texts and totals roughly 242,000 words of running text spread out fairly evenly over the real-time periods sampled in ARCHER. Note that the corpus comes subdivided into seven a priori periods of 50 years.

4.3 The linguistic variable

Recall that this study adopts the variationist methodology. So right at the outset, we circumscribe the variable context to define interchangeable genitive contexts in which either genitive construction is acceptable (Labov 1966, 1972). Using `*' s`, `of`, and `* s` as search strings, Wolk et al. (2013) manually extracted, in a strictly semasiological fashion, all occurrences matching the following patterns:

- [full NP]'s [full NP without determiner];
- [full NP]s [full NP without determiner];
- [full NP]' [full NP without determiner];
- [full NP] *of* [full NP].

Subsequently, Wolk et al. (2013) used a detailed coding scheme to eliminate non-interchangeable genitive contexts (e.g. *of*-constructions with modifying function, or *of* tokens that are part of titles (e.g. *the king of England*)). In this endeavor, Wolk et al. (2013) established criteria on the basis of previous research on genitive variation (e.g. Rosenbach 2002; Hinrichs and Szmrecsanyi 2007). These criteria yield genitive constructions that are interchangeable in principle, rather than necessitating a coder's intuition. The end product is a dataset consisting of N = 3,824 interchangeable genitives covering the period between 1650 and 1999.

4.4 Annotation

Next, the 3,824 genitive observations in the dataset were richly annotated for a range of contextual constraints (or: conditioning factors), including the following:

(i) **Possessor animacy**. According to the literature, this is the most crucial constraint on the genitive alternation: the more human and animate a possessor, the more likely it is to take the *s*-genitive (see, e.g., Altenberg 1982; Rosenbach 2008). The annotation distinguishes between five hand-coded possessor animacy categories (coding scheme: Zaenen et al. 2004):
   1. animate possessors (e.g. *the* ***Bishop's*** *personal security squad* <1979stm2.n8b>)

2. collective possessors (e.g. *the Gentlemen of the **Academy*** <1723dai2.n3b>)
3. time possessors (e.g. *yesterday's outbreaks* <1967stm1.n8b>)
4. locative possessors (e.g. *the inhabitants of this **island*** <1872gla1.n6b>)
5. inanimate possessors (e.g. *the rays of **greatness*** <1748ches.x3b>)

(ii) **Genitive relation**. Genitive constructions can encode a variety of semantic relations. Wolk et al. (2013) follow Rosenbach (2002) and distinguish prototypical and non-prototypical genitive relations. Prototypical relations – which according to the literature favor the *s*-genitive – include:

1. legal ownership relations (e.g. *Mr Ian Smith's cattle ranch and farm* <1979stm1.n8b>)
2. body part relations (e.g. *the murderers legs* <1653merc.n2b>)
3. kinship relations (e.g. *the Duke of Berwick's Son* <1715eve1.n3b>)
4. part-whole relations (e.g. *The Hull of a Ship* <1735rea1.n3b>)

(iii) **Constituent length**. According to the so-called "principle of end-weight", in VO languages such as English speakers and writers tend to place longer, heavier constituents after shorter, lighter ones (for example, Behaghel 1909; Grafmiller and Shih 2011). Thus if the possessor is heavy, there should be a general preference for the *of*-genitive because it places the possessor last; if the possessum is heavy, a general preference for the *s*-genitive is expected. Wolk et al. (2013) operationalized constituent length as a constituent's length in graphemic characters.

(iv) **Final sibilancy**. A final sibilant in the possessor NP (as in *the preparation of this despa**tch***, <1833tim2.n5b>) discourages usage of the *s*-genitive due to a haplology or *horror aequi* effect (for example, Zwicky 1987; Shih et al. 2012). Wolk et al. (2013) annotated all genitives in the dataset as to whether the possessor phrase ends in a sibilant.

The Wolk et al. (2013) paper describes the annotation procedure and the coding schemes on which it is based in much detail. Suffice it to note here that all of the above constraints are non-deterministic in nature – that is, they tend to favor or disfavor particular genitive outcomes in a probabilistic, not categorical, fashion.

4.5 A frequency discussion

In this section, we canvas the frequency distribution of *s*-genitives and *of*-genitives in the Late Modern English period (1650–1999). Table 1 reports raw frequencies (not normalized) as well as genitives rates per ARCHER period (recall that each of these periods covers 50 years). Thus the textual database for the 1650–1699 period (first row) spans about 35,000 words of running text; in this subcorpus, we find in

Table 1: Interchangeable genitive frequencies – absolute (not normalized) versus relative (rates, in %) – by ARCHER period (figures from Wolk et al. 2013).

	of-genitive	*s*-genitive	Total	Corpus size (words)
1650–1699	312 (69%)	139 (31%)	451 (100%)	35k
1700–1749	364 (71%)	152 (29%)	516 (100%)	34k
1750–1799	418 (79%)	109 (21%)	527 (100%)	35k
1800–1849	558 (89%)	70 (11%)	628 (100%)	35k
1850–1899	446 (80%)	109 (20%)	555 (100%)	34k
1900–1949	435 (76%)	134 (24%)	569 (100%)	34k
1950–1999	357 (62%)	221 (38%)	578 (100%)	34k
Total	2,890 (76%)	934 (24%)	3,824 (100%)	242k

all 451 genitive constructions. 312 of these are *of*-genitives, and 139 are *s*-genitives. In terms of relative frequencies, we are therefore talking about an *of*-genitive rate of 69% and an *s*-genitive rate of 31% in the first ARCHER period.

From reading the literature on long-term historical genitive variability – what with the *s*-genitive's comeback during the Early Modern English period and its popularity in Present-Day English (see Section 4.1) – one could have expected to see a gradual linear expansion of *s*-genitive rates during the Late Modern English period. Observe now that no such linear expansion emerges from Table 1. What we find instead is a V-shaped pattern: The *s*-genitive started out with a share of 31% in the 1650–1699 period. Frequencies then started to decline in the 1750–1799 period and reached a low in the 1800–1849 period (11%). Subsequently, *s*-genitive rates recovered such that with a rate of 38% in the 1950–1999 period, the *s*-genitive is more popular now than ever. We note that the V-shaped pattern emerges from relative genitive rates as well as absolute genitive frequencies.

The curious V-shaped frequency pattern that is on display in Table 1 is clearly in need of explanation. Why was the *s*-genitive so unpopular in the nineteenth century? Let us begin by exploring an account that relies on text frequencies as a reliable diagnostic of grammar change (see Szmrecsanyi 2013 for an in-depth discussion). We premise that the *s*-genitive is often discussed as a counterexample to the unidirectionality (less grammatical > more grammatical) of grammaticalization, having developed from a well-behaved inflection in Old English times to a more clitic-like marker in Present-day English. Hence, the history of the English *s*-genitive has been cited as an example of "*de*grammaticalization" (see, e.g., Janda 1980; Newmeyer 1998) or even "*anti*grammaticalization" (Haspelmath 2004). Now, the workhorse diagnostic in the corpus-based grammaticalization literature is a construction's overall text frequency (for example, Krug 2000; Mair 2004): "[l]ack of paradigmatic variability [...] accounts for the ubiquity of a feature in the texts of

a language" (Lehmann 1995: 142). This is why (at least so the argument goes) "increased frequency of a construction over time is prima facie evidence of grammaticalization" (Hopper and Traugott 2003: 129). In this view, we would diagnose stagnation between 1650 and approximately 1850 (which is when *s*-genitive rates collapsed), and re-grammaticalization after 1850, and especially during the twentieth century (which is when *s*-genitive rates recovered substantially).

What is wrong with this account? For starters, these frequency changes in a period of merely 350 years are a bit odd (but then again, strange things do happen). The more severe problem is that the frequencies shown in Table 1 demonstrably entangle language-internal developments (for example, grammaticalization and such like) and language-external developments. Recall from Section 4.4 that animacy of the possessor NP is one of the most crucial conditioning factors in the genitive alternation: animate and/or human NPs favor the *s*-genitive strongly, while inanimate NPs favor the *of*-genitive. And the fact of the matter is that we observe substantial, environmentally induced variability in the distribution of animacy categories (both in terms of genitive NPs and in terms of NPs in general) in written texts during the Late Modern English period. Figure 2 presents two area plots that depict the market share of five animacy categories (*y*-axis, in %) against real time (*x*-axis) in ARCHER's news section. The left plot restricts attention to genitive possessor NPs; the right plot is based on a random sample drawn from the general population of NPs (i.e. not necessarily genitive NPs) in ARCHER news. In both samples, animate nouns are on the decline while

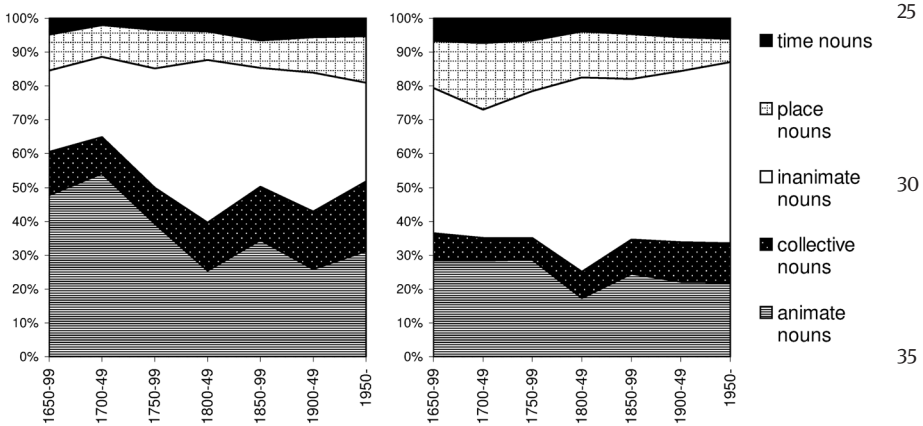

Figure 2: Distribution of animacy categories (*y*-axis, in %) against real time (*x*-axis) in ARCHER news. Left: genitive possessors only. Right: general noun population in ARCHER news (based on a random sample [*N* = 5, 174]).

non-animate nouns are on the rise. In plain English: in the news genre, topics have shifted – from the discussion of animate entities (as in (2a)) to the discussion of non-animate entities, such as collective bodies (as in (2b))

(2) a. *They daily expect here Plenipotentiaries from holland, with a final answer upon* ***the Kings last Propositions*** <1672lon2.n2b>.
b. *Mina had resigned his command, and* ***the orders of the Executive Government*** *were duly obeyed by the local authorities and people of Corunna*. <1822eva1.n5b>

Hence given that the distribution of animacy categories, which boil down to the prime predictor of genitive choice, is unstable in the textual habitat, it should surprise no one that genitive frequencies fluctuate to some extent. The question is: how much of the real-time variance in genitive frequencies can we explain away by considering this habitat instability? To address this issue, we fit a very simple binary logistic regression model (Pampel 2000) that seeks to predict, by dichotomizing predicted probabilities at 0.5, each of the 3,824 genitive outcomes in the dataset *solely on the basis of the animacy status of the possessor*, following the five-fold categorization (animate/collective/time/place/inanimate) presented in Section 4.4.[4] By considering possessor animacy – and possessor animacy only – we aim to regress out unstable animacy distributions from the frequency picture. Given its simplicity, the resulting model has a surprisingly good fit (Somers $D_{xy} = 0.64$) and captures about 40% of the variability in the dataset (Nagelkerke $R^2 = 0.386$). Next, we take the model's 3,824 genitive outcome predictions and calculate from these predictions mean predicted *s*-genitive rates for each of the seven ARCHER periods. These we plot against observed *s*-genitive rates in Figure 3.

Figure 3 shows that the animacy-only model (dotted line) goes some way towards predicting the actually observed collapse of *s*-genitive rates in the nineteenth century (heavy line). Although the animacy-only model admittedly somewhat underestimates the extent of the collapse, it does to some extent account for the frequency decline of the *s*-genitive between 1650 and 1850. This is another way of saying that this part of the story has not much to do with grammar change; what happened – plain and simple – is that a fairly stable grammar of genitive choice produced fewer *s*-genitives because writers chose to write less about animate NPs than they used to. Coming back to the

4 We utilize R (R Development Core Team 2010) package `lrm`, library `Design`.

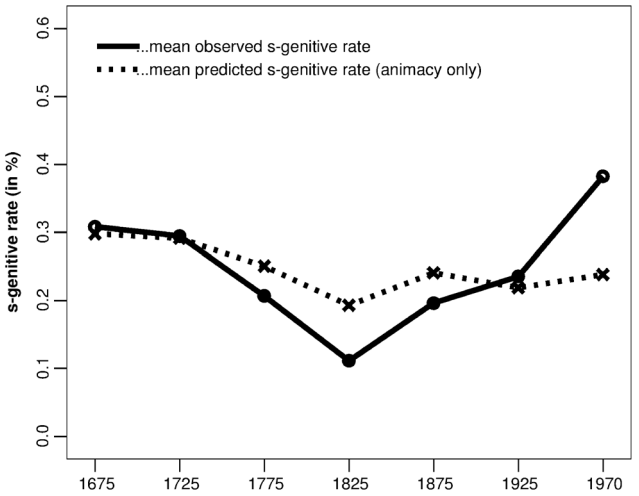

Figure 3: Mean rate of the *s*-genitive (in %, vis-à-vis the *of*-genitive, on *y*-axis) against real time (*x*-axis) (level of granularity: 7 ARCHER periods). Heavy line: observed rates. Dotted line: rates predicted by an animacy-only logistic regression model.

definitions offered in Section 3, we are dealing with a habitat-dependent change that has neither changed the set of basic structural units, not did it necessarily alter the constraint set-up fueling linguistic variability.

That said, we also note that the animacy-only model (that is to say: the habitat-dependence account) completely fails to predict the comeback of the *s*-genitive during the twentieth century. So something else must have happened in addition to the habitat change, and the task before us is to explore what this other development may have been.

4.6 Advanced statistical modeling: Mixed-effects binary logistic regression analysis

We have seen in the previous section that unstable animacy distributions in the textual habitat explain away a good deal of the diachronic variability in genitive outcomes. To judge from the animacy-only model's Nagelkerke R^2 value, about 40% of the frequency variability is trivial. But is the remainder of the frequency variability trivial too? To check if the phenomenal comeback of the *s*-genitive during the twentieth century is a genuine grammar change phenomenon, we will now move on to fit a considerably more sophisticated *mixed-effects logistic*

regression model (Pinheiro and Bates, 2000).[5] The model seeks to predict genitive outcomes in the dataset on the basis of

(i) a variety of language-internal constraints (also known as *fixed effects*), such as the conditioning factors (possessor animacy, genitive relation, constituent length, and final sibilancy) discussed in Section 4.4 above;[6]
(ii) interactions between the language-internal constraints and the language-external variable real time (modeled as a scalar predictor);
(iii) non-repeatable random effects, which control for nuisance factors such as author idiosyncracies and possessor lemma effects.

Wolk et al. (2013) describe the fixed and random effect structure and the model fitting procedure (including bootstrap validation to assess the possibility of overfitting) in ample detail. Suffice it here to spell out our crucial assumption, in line with the definition offered in Section 3: we can posit (probabilistic) grammar change if the stochastic effect of language-internal predictor variables varies as a function of real time. In other words, if we find that the effect of a language-internal constraint is temporally unstable in a statistically significant way, we can legitimately diagnose grammar change (see Gries and Hilpert 2010 for a similar approach). Note that this is a very elegant and precise criterion which injects some, we believe, welcome methodological rigor into the corpus-based study of grammatical change.

The resulting model is a very accurate one: it correctly predicts 91.9% of all genitive outcomes in the dataset, and yields with an excellent Somers D_{xy} value of 0.93. Figure 4 depicts a probabilistic blueprint of genitive choice in ARCHER. The Figure restricts attention to the four conditioning factors discussed in Section 4.4 and reports so-called *odds ratios* (ORs), which quantify the magnitude and the direction of the effect of each predictor on genitive outcomes. ORs specifically indicate how the presence or absence of a feature (for categorical conditioning factors) or how a one-unit increase in a scalar conditioning factor influences the odds for an outcome. Because odds ratios can take values between 0 and infinity, three cases can be distinguished: (i) if OR<1, the conditioning factor makes a specific outcome less likely; (ii) if OR=1, the conditioning factor has no effect whatsoever on the outcome; (iii) if OR>1, the conditioning factor makes a specific outcome more likely (notice that the outcome predicted in Figure 4 is the *of*-genitive). So we observe that all non-animate possessor classes make the *of*-genitive more likely; for example, if the possessor is

5 To fit this model, we utilized the R package `lme4`.
6 In addition, the model considers definiteness of the possessor NP; see Wolk et al. (2013) for details.

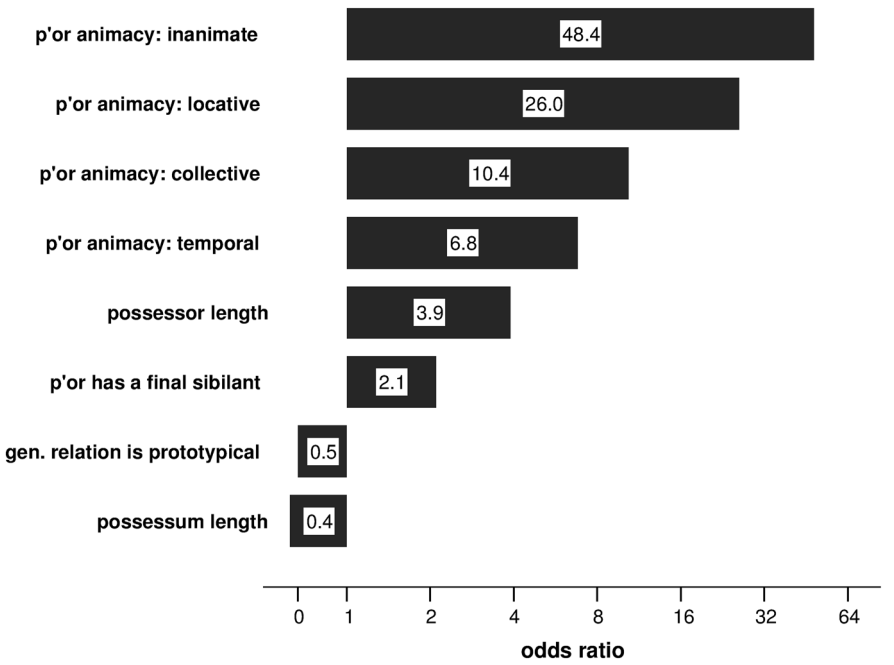

Figure 4: Main effects of genitive predictors: odds ratios (ORs) in logistic regression. Predicted odds are for the *of*-genitive. OR > 1: favoring; OR < 1: disfavoring. Default levels: Animacy – possessor is human; final sibilancy – possessor does not have a final sibilant; genitive relation – non-prototypical (model parameters adapted from Wolk et al. 2013).

inanimate (e.g. *rock*) instead of animate (e.g. *Tom*), the odds for the *of*-genitive increase by a factor of 48. Increasingly long possessors and the presence of a final sibilant in the possessor NP also make the *of*-genitive more likely. By contrast, prototypical genitive relations (e.g. kinship – *Tom's brother*) disfavor the *of*-genitive and hence favor the *s*-genitive: if the genitive relation is prototypical instead of non-prototypical, the odds for the *of*-genitive decrease by 50%. Also, increasingly long possessums favor the *s*-genitive.

These findings are as such nothing to write home about – all of the effect directions are the theoretically expected ones, given the literature. But what about diachronic change? Note now that the blueprint in Figure 4 is an estimate for the year AD 1800. Upon closer inspection, it turns out that the probabilistic effect of some of the conditioning factors depicted interacts significantly with real time.[7] In short, we are indeed dealing with genuine (probabilistic) grammar

7 Hinrichs and Szmrecsanyi (2007) were the first to let language-internal factors interact with real time in regression analysis to diagnose probabilistic language change.

change, according to the criteria that guide this study (see Section 3). Among other things, we obtain a significant interaction between real time and possessor animacy. The technicalities need not concern us here (see Wolk et al. 2013 for a discussion), but what has happened in a nutshell is that starting in the second half of the nineteenth century, the *s*-genitive has been becoming less strongly disfavored with collective, locative and temporal possessors. This is another way of saying that the grammatical animacy constraint has been weakened. For example, whereas in the year 1800, a locative possessor (e.g. *the inhabitants of this island*) increased the odds for an *of*-genitive by a factor of 26 (see Figure 4), the corresponding factor in the year 1999 – two centuries later – is only 6.5. Make no mistake: this odds ratio still robustly favors the *of*-genitive. But the motto in probabilistic grammar is that many a little makes a mickle (especially since collective and temporal possessors have also come to disfavor the *s*-genitive less forcefully), and so a subtle change in the probabilistic constraint set-up fueling genitive variation has engendered a robust frequency change. The result is a substantial frequency boost of the *s*-genitive after 1850.

5 Conclusion

This paper has used statistical data analysis techniques not as an end to themselves, but to distinguish apparent from actual grammatical change. It was argued that this is necessary because the frequency distribution of the target phenomenon and those of its constraints are entangled.

By way of a case study, we have seen that the story of genitive frequencies in the Late Modern English period is complicated. A good deal of the diachronic frequency variability in the dataset can be traced back to environmental changes in the textual habitat (and does not, therefore, diagnose grammar change); but the remainder of the frequency variability is in large part due to (probabilistic) grammar change proper. What happened is that *s*-genitive rates collapsed between 1650 and 1850 because animate NPs became rarer in the textual habitat. So this is the part of the story that has nothing to do with grammar change. However, *s*-genitive frequencies recovered between 1850 and 1999 largely because the grammatical animacy constraint was weakened, such that it became increasingly acceptable to use the *s*-genitive with non-animate possessors. This is the part of the story that indeed involves (probabilistic) grammar change.

The upshot is that frequency shifts do not always reliably diagnose grammar change. To reiterate, text frequencies of the *s*-genitive collapsed in the nineteenth century because news writers, in particular, wrote less and less about

animate entities. This naturally depressed the frequency of the *s*-genitive, a construction which used to be very unpopular with non-animate possessors. It of course remains true that in the long run, environmental changes affecting the textual habitat (such as news writers' increasing interest for non-animate NPs) may very well percolate into grammar (along the lines of Bybee 2006). For example, one may speculate that the frequency collapse of animate possessor NPs in the textual habitat actually triggered the subsequent relaxation of the animacy constraint, via functional adaptation or some such mechanism. But our goal in this paper was not to identify the ultimate causes of grammar change; instead, we sought to establish the extent to which we can trust in text frequencies as a diagnostic marker of grammar change. And in this spirit our case study showed that we can and should differentiate between habitat-induced, environmentally induced frequency fluctuation and grammar change-induced frequency fluctuation. To accomplish this differentiation – which advances historical description and linguistic theory – we suggested going beyond a mere discussion of text frequencies, and exploring instead the probabilistic conditioning of grammatical variability. If language users change the way in which they choose variants, then – and only then – can we explain fluctuating text frequencies as the outcome of grammar change.[8]

So the verdict is that text frequencies are a regrettably unreliable and inconclusive diagnostic of grammar change: they are at best inconclusive, and at worst misleading. The reason is that the all-other-things-being-equal condition, which typically underpins frequency-driven reasoning about grammar change, is rarely met in historical linguistics (and not only in historical linguistics – see Levshina et al. 2013 for similar problems in cross-variety analyses). Thus, we advocate conservatism; before positing grammar change, the analyst needs to rule out alternative explanations. In this context, Occam's razor is useful: the principle reminds us to choose the simplest explanation consistent with the facts – and grammar change, alas, is typically *not* the simplest explanation of frequency fluctuation.

Acknowledgments: I am grateful for the feedback to an earlier version of this paper presented at the 2011 Boston Workshop on "How can new corpus-based techniques advance historical description and linguistic theory?", and for valuable comments and suggestions by two anonymous referees. The usual disclaimers apply. This material is based upon work supported by the National Science

8 One direction for future research would consist of utilizing the Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR) method (Gries and Deshors 2014; Grafmiller 2015) to analyze probabilistic change in diachrony.

Foundation under Grant No. BCS-1025602. This paper is dedicated to the good people at the University Hospital Freiburg (in particular, Prof. Dr. Jürgen Finke, Prof. Dr. Michael Lübbert, Prof. Dr. Hartmut Bertz, and all the staff at the *Station Löhr*), for curing my wife.

References

Altenberg, Bengt. 1982. *The genitive v. the of-construction: A study of syntactic variation in 17th century English*. Malmö: CWK Gleerup.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142.

Biber, Douglas & Susan Conrad. 2001. Register variation: A corpus approach. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The handbook of discourse analysis*, 175–196. Oxford: Blackwell.

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

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86. 186–213.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82. 711–733.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Curzan, Anne. 2009. Historical corpus linguistics and evidence of language change. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 1091–1109. Berlin: De Gruyter Mouton.

Dahl, Lisa. 1971. The s-genitive with non-personal nouns in modern English journalistic style. *Neuphilologische Mitteilungen* 72. 140–172.

Grafmiller, Jason. 2015. Deviant diachrony: Exploring new methods for analyzing language change. Paper presented at "New Developments in the Quantitative Study of Languages", University of Helsinki, August 28–29.

Grafmiller, Jason & Stephanie Shih. 2011. Weighing in on end weight. *Talk given at the LSA 2011 Annual Meeting, 6–9 January 2011, Pittsburgh, Pennsylvania*.

Gries, Stefan Th. & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1). 109–136. doi:10.3366/cor.2014.0053.

Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14. 293–320.

Haspelmath, Martin. 2004. On directionality in language change with particular reference to grammaticalization. In Olga Fischer, Muriel Norde & Harry Perridon (eds.), *Up and down the cline*: *The nature of grammaticalization*, 17–44. Amsterdam: John Benjamins.

Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11. 437–474.

Hopper, Paul J. & Elizabeth Closs Traugott. 2003. *Grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.
Hundt, Marianne & Christian Mair. 1999. 'Agile' and 'uptight' genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4. 221–242.
Janda, Richard D. 1980. On the decline of declensional systems: The overall loss of OE nominal case inflections and the ME reanalysis of *-es* as *his*. In Elizabeth Closs Traugott, Rebecca Labrum & Susan Shepherd (eds.), *Papers from the 4th International Conference on Historical Linguistics*, 243–252. Amsterdam: John Benjamins.
Jucker, Andreas. 1993. The genitive versus the of-construction in newspaper language. In Andreas Jucker (ed.), *The noun phrase in English: Its structure and variability*, 121–136. Heidelberg: Carl Winter.
Krug, Manfred G. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: Mouton de Gruyter.
Labov, William. 1966. The linguistic variable as a structural unit. *Washington Linguistics Review* 3. 4–22.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. 1982. Building on empirical foundations. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Perspectives on historical linguistics*, 17–92. Amsterdam: John Benjamins.
Lavandera, Beatriz R. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7(2): 171–182.
Lehmann, Christian. 1995. *Thoughts on grammaticalization*. München: LINCOM EUROPA.
Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013. Towards a 3D-Grammar: Variation in the use of Dutch causative constructions. *Journal of Pragmatics* 52. 34–48.
Mair, Christian. 2004. Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In Christian Mair & Hans Lindquist (eds.), *Corpus approaches to grammaticalisation in English*, 121–150. Amsterdam: John Benjamins.
Mustanoja, Tauno F. 1960. *A Middle English syntax*, vol. I. Helsinki: Société Néophilologique.
Newmeyer, Frederick J. 1998. *Language form and language function*. Cambridge, MA: MIT Press.
Pampel, Fred. 2000. *Logistic regression. A primer*. Thousand Oaks: Sage Publications.
Paul, Hermann. 1920. *Prinzipien der Sprachgeschichte*, 5th edn. Halle: Niemeyer.
Pinheiro, José C. & Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.
Potter, Simeon. 1969. *Changing English*. London: André Deutsch.
R Development Core Team. 2010. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07–0.
Raab-Fischer, Roswitha. 1995. Löst der Genitiv die of-Phrase ab? Eine korpusgestützte Studie zum Sprachwandel im heutigen Englisch. *Zeitschrift für Anglistik und Amerikanistik* 43(2). 123–132.
Raumolin-Brunberg, Helena. 2005. The diffusion of subject YOU: A case study in historical sociolinguistics. *Language Variation and Change* 17. 55–73.
Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin: Mouton de Gruyter.
Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 379–412. Berlin: Mouton de Gruyter.

Rosenbach, Anette. 2008. Animacy and grammatical variation – findings from English genitive variation. *Lingua* 118(2). 151–171.
Sapir, Edward. 2004. *Language: An introduction to the study of speech*. Mineola, NY: Dover.
Shih, Stephanie, Jason Grafmiller, Richard Futrell & Joan Bresnan. 2012. Rhythm's role in genitive construction choice in spoken English. In Ralf Vogel & Ruben van de Vijver. (eds.), *Rhythm in phonetics, grammar and cognition*. Berlin & Boston: de Gruyter.
Szmrecsanyi, Benedikt. 2009. Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21(3). 319–353.
Szmrecsanyi, Benedikt. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in cognitive sociolinguistics*, 139–166. Berlin: De Gruyter Mouton.
Szmrecsanyi, Benedikt. 2013. The great regression: Genitive variability in Late Modern English news texts. In Kersti Börjars, David Denison & and Alan K. Scott (eds.), *Morphosyntactic categories and the expression of possession*, 59–88. Amsterdam: John Benjamins.
Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: John Benjamins.
Tagliamonte, Sali. 2001. Comparative sociolinguistics. In J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 729–763. Oxford: Blackwell.
Thomas, Russell. 1931. *Syntactical processes involved in the development of the adnominal periphrastic genitive in the English language*. PhD thesis, University of Michigan.
Weiner, Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19: 29–58.
Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.
Yánez-Bouza, Nuria. 2011. ARCHER past and present: 1990–2010. *ICAME Journal* 35: 205–236.
Zaenen, Annie, Jean Carlette, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor & Tom Wasow. 2004. Animacy encoding in English: Why and how. In Donna Byron & Bonnie Webber (eds.), *Proceedings of the 2004 ACL workshop on discourse annotation, Barcelona, July 2004*, 118–125. East Stroudsburg, PA: Association for Computational Linguistics (ACL).
Zwicky, Arnold M. 1987. Suppressing the zs. *Journal of Linguistics* 23. 133–148.