Diachronic Probabilistic Grammar

Benedikt Szmrecsanyi (KU Leuven)

Acknowledgments

I am grateful to Hubert Cuyckens and Frauke D'hoedt for helpful comments and suggestions. The usual disclaimers apply.

Abstract

The paper sketches a novel, usage-based framework – Diachronic Probabilistic Grammar (DPG) – to analyze variation and change in diachrony. The approach builds on previous work in the Probabilistic Grammar tradition (see, for example, Bresnan 2007; Bresnan and Ford 2010) demonstrating, based on converging experimental and observational evidence, that syntactic knowledge is to some extent probabilistic, and that language users have excellent predictive abilities. What takes center stage in the approach is how contextual predictors (such as, for example, the principle of end weight) constrain linguistic variation. DPG is specifically interested in the extent to which such probabilistic constraints are (un)stable in the course of time. To highlight the diagnostic potential of the DPG framework, the paper explores three case studies: the development of the alternation between non-finite and finite complementation in the Late Modern English period, recent changes in the genitive alternation in the late 20th century, and a cross-constructional analysis of parallelisms in the development of the genitive and the dative alternation in the Late Modern English period.

Key words: historical linguistics, probabilistic grammar, corpus linguistics, dative alternation, genitive alternation, complementation

1. Introduction

This paper discusses a new approach to studying syntactic variation in diachrony. Methodologically, we draw on richly annotated datasets and statistical modeling to explore the development of probabilistic constraints on syntactic variation, fully concurring with Adger and Trousdale (2007, 274) that variation is a "core explanandum" in linguistics. On the theoretical plane, we use the Probabilistic Grammar framework (in the spirit of, e.g., Bresnan 2007; Bresnan and Ford 2010) to interpret the historical evolution of three syntactic alternations in the grammar of English: the alternation between non-finite and finite complementation, as in (1); the alternation between the *s*-genitive and the *of*-genitive, as in (2); and the alternation between the ditransitive dative and the prepositional dative, as in (3).

(1) a. I don't regret helping her start out
(non-finite complementation)
b. I don't regret that I helped her start out
(finite complementation)

(2) a. The Seneschal's brother
(the *s*-genitive)
b. The brother of the Seneschal
(the *of*-genitive)

(3) a. wrote M. an earnest loving note
(the ditransitive dative)
b. wrote a note to M.
(the prepositional dative)

Methodologically, the Diachronic Probabilistic Grammar (DPG) approach consists of five steps: (1) tap into text corpora (that is, observational usage data) and use the variationist methodology (e.g. Labov 1982) to derive richly annotated datasets; (2) fit statistical models that predict language users' syntactic choices from language-internal predictors (also known as "conditioning factors" or "constraints"); (3) explore real-time changes in the effect that these predictors have; (4) interpret any such changes in terms of diachronically evolving probabilistic grammars; (5) interpret the absence of changes as probabilistic stability.

The statistical analysis technique that will take center stage in this paper is binary logistic regression analysis. The workhorse analysis technique in corpus-based variation studies, logistic regression probes the probabilistic conditioning – and its plasticity in real time – of linguistic choice-making. The technique predicts a binary outcome (i.e. a linguistic choice) given a range of independent predictor variables. Thanks to multivariate control, regression analysis is the closest historical linguists can come to conducting a controlled experiment. Regression analysis is increasingly popular in corpus-based historical linguistics (see, e.g., Gries and Hilpert 2010; Hundt and Szmrecsanyi 2012; Wolk et al. 2013), and state-of the-art designs factor in real time by fitting interaction effects between time as a language-external variable and various language-internal predictors. These interaction effects can gauge if and to what extent the probabilistic effects of language-internal predictors are stable or unstable diachronically.

This paper is structured as follows: in Section 2, we discuss in more detail the theoretical underpinnings of the paper. In Section 3, we present the three empirical case studies: complementation strategy choice in the Late Modern English period (Section 3.1), genitive variability in the late 20th century (Section 3.2), and dative and genitive variability in the Late Modern English period (Section 3.3). Section 4 offers a discussion and some concluding remarks.

2. Theory and background

In most general terms, the approach outlined in this paper is an exercise in probabilistic linguistics (see Bod, Hay, and Jannedy 2003 for papers in this spirit). The analysis will specifically rely on the variation-centered, usage- and experience-based Probabilistic Grammar framework developed by Joan Bresnan and collaborators (Bresnan 2007; Bresnan et al. 2007; Bresnan and Hay 2008; Bresnan and Ford 2010; Wolk et al. 2013). The framework makes three crucial assumptions, which are broadly compatible with modern variationist sociolinguistic theory (Labov 1982; Tagliamonte 2001):

Assumption (1): Grammatical variation is sensitive not (only) to categorical constraints, but to multiple and typically conflicting probabilistic constraints, be they formal, semantic, or phonological in nature. Such constraints, like the principle of end-weight (place longer constituents after shorter constituents), may influence linguistic choice-making in subtle ways (Bresnan and Hay 2008).

Assumption (2): Linguistic knowledge includes knowledge of probabilities, and language users have powerful predictive capacities (Gahl and Garnsey 2006; Gahl and Yu 2006).

Assumption (3): Corpus-based regression models match speakers predictive abilities (Bresnan 2007; Bresnan and Ford 2010).

How do we know that these assumptions are true? Assumption (1) is fairly uncontroversial in the usage-based and empirical linguistics literature. As to assumptions (2) and (3), Bresnan and collaborators have shown in a series of experiments that the likelihood with which we find particular syntactic variants in a corpus corresponds to the intuitions that native speakers have about the naturalness of these variants, given the same context. Bresnan (2007), for example, is interested in the probabilistic underpinnings of the dative alternation (*I sent the president a letter* versus *I sent a letter to the president*), and so she departs from the dative regression model presented in Bresnan et al. (2007). This model predicts dative choices in the Switchboard corpus of spoken English with 94% accuracy, given a range of language-internal predictors (e.g. length of the constituents, information status, animacy of the constituents, and so on). Note now that regression models such as the one reported in Bresnan et al. (2007) do not only predict categorical choices but actually assign a probability (e.g. of usage of the prepositional dative construction, as in *I sent a letter to the president*, instead of a ditransitive dative construction) to each and every observed dative occurrence (prepositional or not) in the dataset. Having thus at her disposal a corpus-based regression model that calculates realization probabilities based on usage data, Bresnan (2007) subsequently moves on to a type of experiment that is now known as the “100-split” task. Ford and Bresnan (2013, PAGE TO BE INSERTED) summarize this task as follows:

> Participants rate the naturalness of alternative forms as continuations of a context by distributing 100 points between the alternatives. Thus, for example, participants might give pairs of values to the alternatives like 25-75, 0-100, or 36-64. From such values, one can determine whether the participants give responses in line with the probabilities given by the model and whether people are influenced by the predictors in the same manner as the model.

In short, participants are confronted with the same sort of material (prepositional or ditransitive dative constructions in context) that the corpus-based regression model had been confronted with. It turns out that there is a significant correlation between participants’ ratings and the probabilities calculated by the regression model. Thus, matching up corpus and experimental methods shows that “language users can in effect make accurate probabilistic predictions of the syntactic choices of others” (Bresnan 2007, 91), and that regression modeling captures aspects of language users’ linguistic knowledge.

The task before us in this paper is to transfer Probabilistic Grammar framework to the realm of historical data. The methodological challenge, of course, is that past speakers/writers (from the 17th century, say) are not available for experimental testing; all we have is text corpora sampling these individuals’ production data in written form. But we can still apply the uniformitarian principle, and assume that the cognitive mechanisms underlying present-day probabilistic patterns also underlie past variation (see Jäger and Rosenbach 2008): if present-day language users’ linguistic knowledge includes knowledge about probabilities, so did past language users’ linguistic knowledge; if present-day language users have powerful predictive abilities, so did 17th century language users. The Diachronic Probabilistic Grammar frameworkthus endeavors to model past speakers'/writers' implicit grammatical knowledge, based on observational data. This is another way of saying that DPG is *not* merely interested in describing variation in corpus data drawing on the mathematics of uncertainty. Instead, DPG ultimately aims to explore the extent to which constraints on syntactic variation – and knowledge of them – are historically (un)stable.

Let us summarize the foregoing discussion. Linguistic knowledge includes knowledge about probabilities, and we can use corpus data to model this knowledge. Against this backdrop, the primary objective of the DPG approach is to explore the extent to which probabilistic knowledge about grammar evolves over time. The primary empirical diagnostic of probabilistic change is the existence of robust interaction effects between language-internal predictors (such as weight effects) and real time in regression analysis. Thus, if a regression model finds, for example, that in the 18th century each additional word in a genitive possessor phrase increased the odds for an *of*-genitive, say, twofold while in the 20th century each additional word increased the odds by a factor of, say, three, we are dealing with a probabilistic grammar change that can be interpreted as a diachronic change in probabilistic knowledge.

3. Case studies

In this section, we discuss three case studies to demonstrate how DPG can inform our interpretation of syntactic variation and change. Because we have published focused papers on each of the alternations that will be discussed in the following sections, the descriptions of the technicalities will be kept to a minimum; interested readers are referred to the original papers for details.

3.1. Late Modern English complementation

Our first case study is concerned with complementation in the Late Modern English period. Specifically, we will explore if the probabilistic factors that constrain the choice between finite and non-finite complementation have been stable in the Late Modern English period or not. Complementation has been an important research topic in the generative as well as cognitive-functional literature (see, for example, Bresnan 1970; Givón 1980). Cuyckens, D'hoedt, and Szmrecsanyi (to appear) embark on a probabilistic analysis of historical complement-clause (CC) variation with the complement-taking predicates (CTPs) *remember*, *regret*, and *deny* (based on a list of factual verbs which exhibit the finite/non-finite CC alternation, according to Quirk et al. 1985, 1182–1184). It is this study that we take the liberty to summarize and re-interpret in what follows.

Cuyckens, D'hoedt, and Szmrecsanyi (to appear) are interested, for one thing, in finite complement clauses which are introduced either by the complementizer *that* (4a) or by zero (4b).

(4) a. I remember$_{CTP}$ perfectly well [*that* it was at the prisoner's suggestion] $_{CC}$
<Old Bailey Corpus, t-18720226–267>
(finite complement clause introduced by *that*)
b. I remember$_{CTP}$ [___ Boswell and Ausser were both at my house] $_{CC}$
<Old Bailey Corpus, t17670715–45>
(finite complement clause introduced by zero)

In many cases, such finite structures vary with non-finite patterns: subjectless *–ing* CCs, as in (5a), *–ing* CCs with expressed subject, as in (5b), subjectless *to*-infinitive CCs, as in (5c), and *to*-infinitive CCs with expressed subjects, as in (5d).

(5) a. Do you remember$_{CTP}$ [at any time go*ing* to the prisoner's house] $_{CC}$?
<Old Bailey Corpus, t17730217–52>
(subjectless *–ing* CC)
b. Do you remember$_{CTP}$ [*a green cart* com*ing* up] $_{CC}$?

<Old Bailey Corpus, t17970920–62>
(*–ing* CC with expressed subject)

c. I do not remember$_{CTP}$ [ever *to have heard* a word from you before]$_{CC}$
<Corpus of Late Modern English Texts, George Byron, Letters>
(subjectless *to*-infinitive CC)

d. I do remember$_{CTP}$ [*this circumstance to have happened* but to one man]$_{CC}$
<Old Bailey Corpus, t17950218–46>
(*to*-infinitive CC with expressed subject)

Cuyckens, D'hoedt, and Szmrecsanyi (to appear) explored the above patterns in two corpora covering the period between 1710 and 1920: the *Old Bailey Corpus* (OBC) version 0.9 (Huber et al. 2012), which samples court transcripts and is thus relatively close to the spoken language; and the *Corpus of Late Modern English Texts*, extended version (CLMETEV) (see https://perswww.kuleuven.be/~u0044428/), which mainly contains fictional texts. From this database, Cuyckens, D'hoedt, and Szmrecsanyi (to appear) extracted all occurrences of *remember*, *regret*, and *deny* followed by a CC. The dependent variable in the study is thus CC type: non-finite clauses (as in (5)) versus finite clauses (as in (4)). The resulting dataset comprises *N* = 5,228 CC occurrences (*remember*: 3,810 observations; *regret*: 280 observations; and *deny*: 1,138 observations).

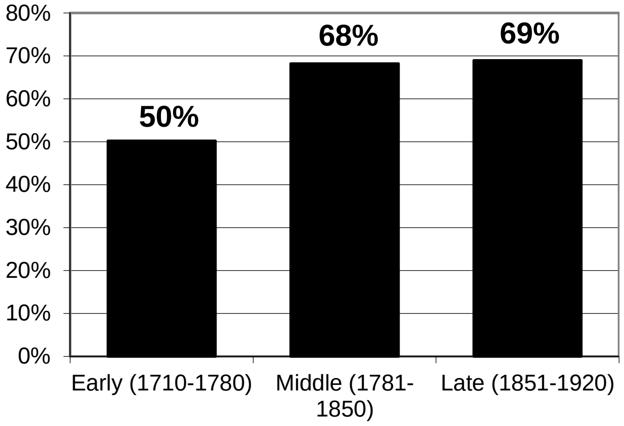

Figure 1. Rates of non-finite complementation (*y*-axis) by real time period (*x*-axis). The difference between the early and the two later periods is significant at *p* < .001.

We begin by canvassing the rates of non-finite complementation, vis-à-vis finite complementation, in real time. Figure 1 splits up the Late Modern English period into three sub-periods. It is amply clear from the Figure that non-finite complementation has been on the rise in the Late Modern English period: in the Early period (1710-1780), non-finite complementation had a market share of 50%. In the Middle (1781-1850) and the Late period (1851-1920), the share of non-finite complementation amounts to 68% and 69%, respectively.

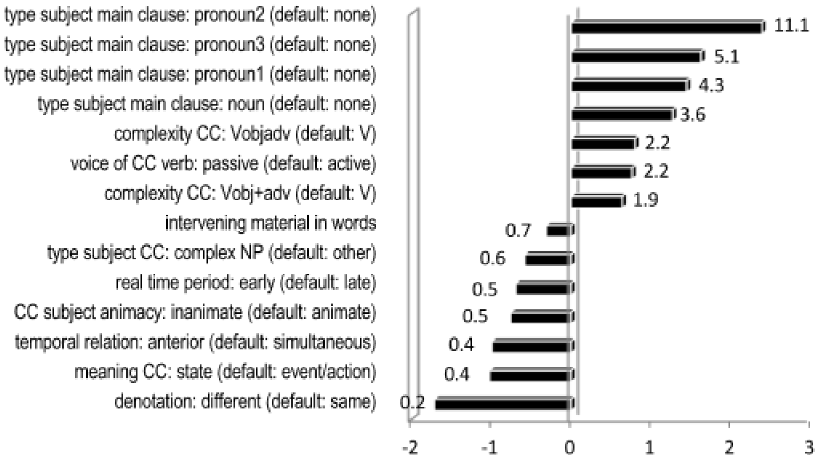

Figure 2. Choice of complementation strategy in the Late Modern English period: significant main effects in a minimal adequate mixed-effects logistic regression model. Bars plot regression coefficients (> 0: favoring, < 0: disfavoring); figures indicate odds ratios (> 1: favoring, < 1: disfavoring). Predicted odds are for non-finite complementation. *N* = 5,228; correctly predicted: 84% (baseline: 64%); Somers Dxy = .81, κ = 21.7. Random effects: medium, verb meaning, corpus file ID.

To identify the determinants of the variability, Cuyckens, D'hoedt, and Szmrecsanyi (to appear) coded the dataset for various pertinent language internal-factors, for example

- **type of the subject in the main clause**: noun, 1st/2nd/3rd person pronoun, none
- **meaning of the CC**: state, event/action
- **type of the subject in the CC**: complex NP, other
- **CC subject animacy**: animate, inanimate
- **voice of the verb in the CC**: active, passive, copular
- **intervening material (in words) between complement-taking predicate and CC**
- **denotation**: same (main clause subject and CC subject denote same entities) versus different
- **temporal relation**: anterior, posterior, simultaneous

In addition, of course, Cuyckens, D'hoedt, and Szmrecsanyi (to appear) recorded the time period from which each observation derives. On the basis of this annotation (which we will draw on to investigate if the rise of non-finite complementation is due to probabilistic changes in the effect that various constraints have on the variability), Cuyckens, D'hoedt, and Szmrecsanyi (to appear) fit a regression model that predicts the odds for non-finite complementation. In Figure 2, we find the sketch of a regression model – and thus, a probabilistic footprint – of complementation strategy choice in their dataset. Thus, for example, we learn that if the subject of the main clause is a 2nd person pronoun (first row in Figure 2), the odds for non-finite complementation increase by a factor of 11.1; if the temporal relation between the CC and the main clause is one of anteriority (as in *I regret that I helped her out*) (row 12 in Figure 1), the odds for non-finite complementation decrease by a factor of 0.4, that is, by 60%.

Let us now scrutinize the role that real time plays in the regression model. Remember that our primary empirical diagnostic of probabilistic change is robust interaction effects between language-internal predictors and real time in regression analysis. Cuyckens, D'hoedt, and Szmrecsanyi (to appear) indeed find three relatively minor interactions between language-internal predictors and real time; for example, the anteriority effect is becoming a bit weaker in the course of time (see Cuyckens, D'hoedt, and Szmrecsanyi to appear for details). But note that the model hardly suffers when these interactions are removed (predictive accuracy hardly decreases, from 84.3% to 84.1%). Instead, real time has a substantial main effect (*p* < .01) in the dataset: in the early period, regardless of the linguistic context, the odds for non-finite complementation are only half as big as in the late period (see Figure 2, row 10).

In summary, then, the alternation between finite and non-finite complementation is probabilistically fairly stable in the Late Modern English period, in that real time does not interact much with language-internal predictors – it just so happens that non-finite complementation is overall becoming more frequent in the course of time. This frequency increase of non-finite complementation in the period under analysis (see Figure 1) is a likely outgrowth of the fact that "a long-term trend in English has been the growth of nonfinite complement clauses at the expense of finite clauses" (Denison 1998, 256). In short, we are dealing with a generic, not alternation-specific drift that does not come within the remit of probabilistic change. Therefore, the DPG framework as defined in the present paper does not diagnose probabilistic change in the dataset under analysis.

3.2. 20th century genitive variability

In the previous case study we did not see any substantial probabilistic change – to set the scene this paper reported a null finding, as it were, to demonstrate that DPG is not an "anything goes" approach. Things are different, though, with regard to genitive variability in the second half of the twentieth century. Consider Hinrichs and Szmrecsanyi (2007), who explore the alternation between the s-genitive, as in (6a), and the of-genitive, as in (6b).

(6) a. [The bill]$_{possessor}$'s [supporters]$_{possessum}$ said they still expected Senate approval of the complex and sweeping energy package
<Frown, A02>
(the *s*-genitive)

b. Latter domain, under the [guidance]$_{possessum}$ of [Chef Tom Yokel]$_{possessor}$, will specialize in steaks, chops, chicken and prime beef
<Brown, A31>
(the *s*-genitive)

Hinrichs and Szmrecsanyi (2007) specifically explore recent (late 20th century) probabilistic changes in the genitive alternation, based on the "Reportage" and "Editorial" genres in the original "Brown family" of corpora, which consists of four 1 million-word corpora with (near-)identical design. The corpora represent written language, drawn from 1960s American English (the Brown corpus), 1960s British English (the LOB corpus), 1990s American English (the Frown corpus), and 1990s British English (the LOB corpus). Figure 3 visually depicts the design of the Brown family of corpora (see Hinrichs, Smith, and Waibel 2010 for the corpus manual).

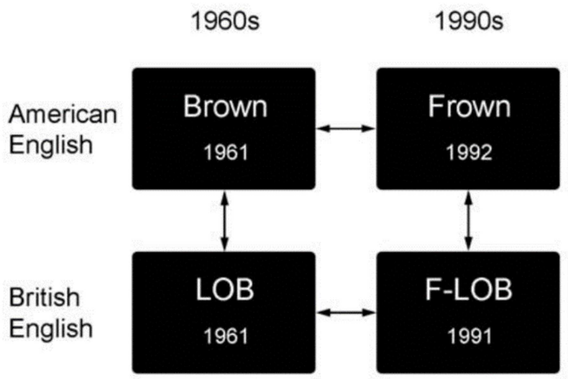

Figure 3. The original Brown family of corpora.

Among other things, Hinrichs and Szmrecsanyi (2007)report that the *s*-genitive is on the rise overall during the second half of the 20th century because it is the more economical coding option, and that *s*-genitive is more frequent in American English than in British English because it is less constrained by the animacy constraint in American English. In this section, we endeavor to re-interpret the study against the backdrop of the DPG framework.

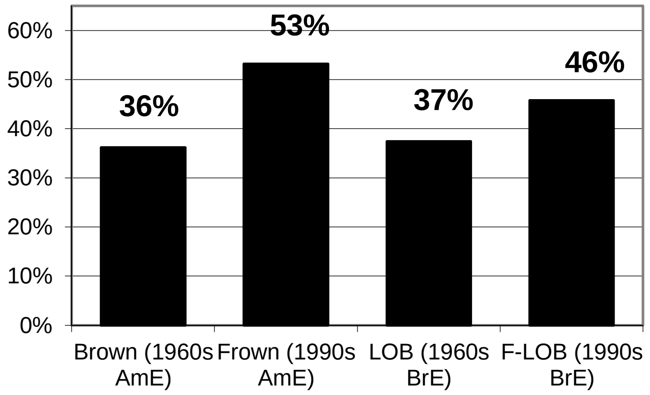

Figure 4. Rates of the *s*-genitive in the Brown family of corpora.

Hinrichs and Szmrecsanyi (2007) identified potential genitive occurrences in the data (i.e. strings with *’s* and *of*), and handcoded these for interchangeability. In other words, only those genitives were retained that could be paraphrased by the alternative construction (see Hinrichs and Szmrecsanyi 2007, 445–447 for a detailed coding scheme); the genitive in (6) qualify as interchangeable because (6a) could be paraphrased as *the supporters of the bill* and (6b) as *Chef Tom Yokel’s guidance*. The crucial dependent variable in Hinrichs and Szmrecsanyi (2007) is genitive realization (*‘s* versus *of*), and their dataset spans N = 8,300 interchangeable genitive observations. Figure 4 plots s-genitive rates across the four components of the Brown family. Observe that while in the 1960s, the *s*-genitive was used 36%-37% of the time (Brown and LOB), its market share in the 1990s is much higher: 53% in Frown and 46% in F-LOB. So the issue we will be investigating in what follows is if and to what extent probabilistic grammar change to blame for the rise of the *s*-genitive in late 20th century written English.

In addressing this question, the rich contextual annotation of the dataset explored in Hinrichs and Szmrecsanyi (2007) is helpful. Hinrichs and Szmrecsanyi annotated the interchangeable genitive observations in their dataset for a range of conditioning factors known to constrain genitive variability, including

- **animacy of the possessor**: human, animal, collective, inanimate (it is well-known that animate possessors attract the *s*-genitive)
- **presence of a nested of -genitive**, as in [*the boss*] *of* [[*the father*] *of* [*the bride*]]
- **presence of a nested s-genitive** [*the boss*] *of* [[*the bride*]'s [*father*]]
- **type-token ratio of embedding passage**, to measure lexical density
- **text frequency of possessor head**, to measure thematicity of the possessor
- **persistence**, a.k.a. syntactic priming
- **nouniness of embedding passage**
- **final sibilancy in the possessor**, as in *Alice's brother*
- **possessor length** (in words): the principle of end weight

Needless to say, each genitive observation was also annotated for its real time period (1960s written English versus 1990s written English). Subsequently, Hinrichs and Szmrecsanyi fit a binary logistic regression model, to quantify the probabilistic impact of the above conditioning factors on genitive choice, and to check if the probabilistic effect of these factors was subject to change between the 1960s and the 1990s.

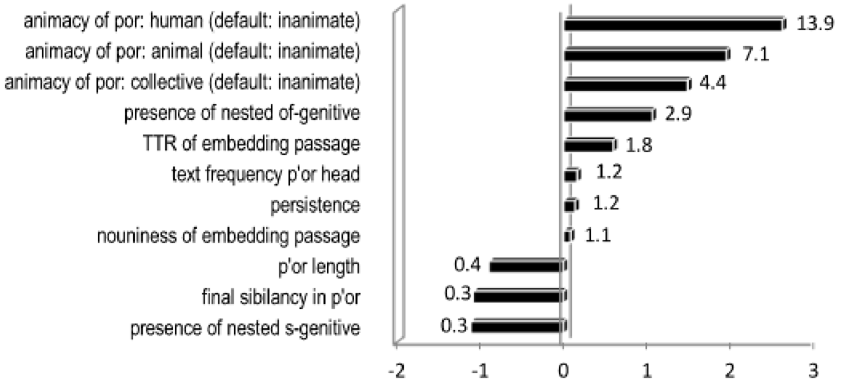

Figure 5. Genitive choice in late 20th century written English: significant main effects in a minimal adequate fixed-effects logistic regression model. Bars plot regression coefficients (> 0: favoring, < 0: disfavoring); figures indicate odds ratios (> 1: favoring, < 1: disfavoring). Predicted odds are for the *s*-genitive. *N* = 8,015; correctly predicted: 79% (baseline: 57%); Nagelkerke R^2 = 0.51.

A sketch of the Brown family regression model of genitive choice reported in Hinrichs and Szmrecsanyi (2007) is provided in Figure 5. The effect direction of the constraints modeled is the theoretically expected one, given the literature. For example, vis-à-vis inanimate possessors (as in *the consequences of inflation*), human possessors attract the *s*-genitive (as they should, given the literature -- see, e.g., Rosenbach 2005): if the possessor is human instead of inanimate, the odds for an *s*-genitive realization increase by a factor of 13.9. Conversely, thanks to the principle of end-weight (e.g. Arnold et al. 2000; Behaghel 1909), every additional word in the possessor phrase decreases the odds for an *s*-genitive realization by a factor of 0.4. Long possessors attract the *of*-genitive because the *of*-genitive places the possessor after the possessum.

We now turn our attention to the role that real time plays in the model. Observe, first, that unlike in complementation strategy choice (Section 3.1), real time does *not* have a significant main effect in the regression model reported in Hinrichs and Szmrecsanyi (2007). In other words, there is no evidence for a "magical drift" towards the *s*-genitive. Unlike non-finite complementation after *remember, regret,* and *deny* – which is part and parcel of a more comprehensive drift in the English towards non-finite verb forms – the *s*-genitive does not appear to surf a bigger tide.

But real time does come into the picture in another way. Recall again that our primary empirical diagnostic of probabilistic change is robust interaction effects between language-internal predictors and real time in regression analysis. And indeed Hinrichs and Szmrecsanyi (2007) uncover three robust and statistically significant interaction effects between language-internal predictors and real time:

1. **Text frequency of possessor head:** Osselton (1988) pointed out that under normal circumstances, an inanimate noun such as *inflation* is unlikely to take the *s*-genitive. Yet in an economics textbook, we may very well find the phrase *inflation's consequences*. In other words, we are dealing with a frequency effect such that highly thematic nouns are more likely to take the *s*-genitive, all other things being equal. Hinrichs and Szmrecsanyi (2007) measured thematicity by determining the natural logarithm (*ln*) of the possessor head noun in the corpus text in which the genitive is observed. Their model shows that while in the 1960s, the odds for the s-genitive increase by a factor of 1.18 for every one-unit increase in the measure, the corresponding factor in the 1990s comes out as 1.65. Therefore, 1990s writers are more sensitive to the thematicity of the possessor head noun than 1960s writers.
2. **Final sibilancy in the possessor:** A final sibilant in the possessor NP is claimed to encourage usage of the of genitive due to a haplology or horror aequi effect (e.g. Altenberg 1982; Zwicky 1987); Hinrichs and Szmrecsanyi (2007) used a script to automatically annotate possessors for final sibilancy, based on orthographic spelling. In their 1960s data, a final sibilant discourages usage of an *s*-genitive with an odds ratio of .34; in their 1990s data, the effect is stronger, having an odds ratio of 0.25. Thus, compared to 1960s writers, 1990s writers are more sensitive to the phonological context.
3. **Possessor length:** we already saw that longer possessors disfavor the *s*-genitive (and favor the *of*-genitive), thanks to the principle of end weight. As it turns out, the probabilistic effect of possessor length is more disfavoring in the 1990s (odds ratio: 0.27) than in the 1960s (odds ratio: 0.41). Cross-variety comparison shows that it is mainly British writers (rather than American writers) who have come to pay more attention to the length of the possessor phrase when choosing genitives.

By way of an interim summary, we have seen in this section that late 20th century genitive grammars were subject to probabilistic change, in that the magnitude of the effect of some constraints is variable in real time: writers have come to assign different probabilities to certain contexts when choosing between *s*- and *of*-genitives. More specifically, we have seen that journalists increasingly use the *s*-genitive when the possessor is highly thematic, and that they increasingly disfavor the *s*-genitive when the possessor ends in a final sibilant, or when it is long. It is important to note that our claim is not that writers' categorical knowledge about the genitive alternation (e.g. what constitutes an acceptable *s*-genitive) has changed. Rather, drawing on DPG diagnostics we claim that 1960s writers' probabilistic knowledge about genitive grammars subtly differs from that of 1990s writers, in that 1990s writers are more or less likely to use particular genitive variants in particular contexts.

3.3. Late Modern English genitive and dative variability

The third case study to be presented in this paper takes a joint look at dative and genitive variability in the Late Modern English period. This section draws on empirical findings and interpretations originally presented in Wolk, Bresnan, Rosenbach, and Szmrecsanyi (2013), who present a cross-constructional probabilistic analysis of the history of the genitive alternation (as in (2), reproduced below as (7)) and the dative alternation (as in (3), reproduced bellow as (8)) in the Late Modern English period.

(7) a. before [*The Seneschal*]$_{possessor}$*'s* [*brother*]$_{possessum}$ could arrive, he was secured by the Governor of Newport
<ARCHER, 1682pro1.n2b>
(the *s*-genitive)

b. the Duke of Norfolk, having lately received another Challenge from [*the brother*]$_{possessum}$ *of* [*the Seneschal*]$_{possessor}$, went to the place appointed
<ARCHER, 1682pro1.n2b >
(the *of*-genitive)

(8) a. SUN., JAN. 23 — M.'s birthday — *wrote* [*M.*]$_{recipient}$ [*an earnest loving note*]$_{theme}$
<ARCHER, 1887gibs.j6a>
(the ditransitive dative)

b. *wrote* [*a note*]$_{theme}$ *to* [*M.*]$_{recipient}$ expressive of my good state of feeling.
<ARCHER, 1887gibs.j6a>
(the prepositional dative)

Why adopt a cross-constructional perspective to study the genitive and dative alternation? Notice, first, that the two alternations exhibit a number of similarities: there are distributional similarities (the generally similar probabilistic constraints on realization choice, such as the principle of end weight). Second, there are formal similarities: both alternations are essentially word order alternations, where the order of the possessor/recipient and possessum/theme can be manipulated. Third, the two alternations share a common core of meaning ("[potential] possession"). Genitive and dative variability thus offers an exciting target for cross-constructional analysis.

Wolk at al. (2013) tap into A Representative Corpus of Historical English Registers, release 3.1 (ARCHER) (Yánez-Bouza 2011). ARCHER covers the period between 1650 and 1999, spans about 1.8 million words of running text, and samples eight different registers (drama, fiction, sermons, journals/diaries, medicine, news, science, letters) and the two major varieties of English, British and American (coverage of American English is restricted to three of the seven periods, however). Wolk et al.'s investigation of the dative alternation draws on the ARCHER corpus in its entirety (that is, all periods, registers, and both American and British texts). Genitives are substantially more frequent than datives, which is why attention is restricted to alternating genitives in ARCHER's British English news and letters sections. From the corpus material, Wolk at al. (2013) extracted interchangeable *s*- and *of*-genitives roughly following the guidelines in Hinrichs and Szmrecsanyi (2007) (see Section 3.2). Interchangeable ditransitive and prepositional dative occurrences were identified by first defining a list of verbs that can appear with a dative object. Subsequently, ARCHER was searched for instances of these verbs that were followed by two NP argument slots. Non-interchangeable datives and other constructions (e.g. benefactive ditransitives, as in *make us some tea*) were ignored.

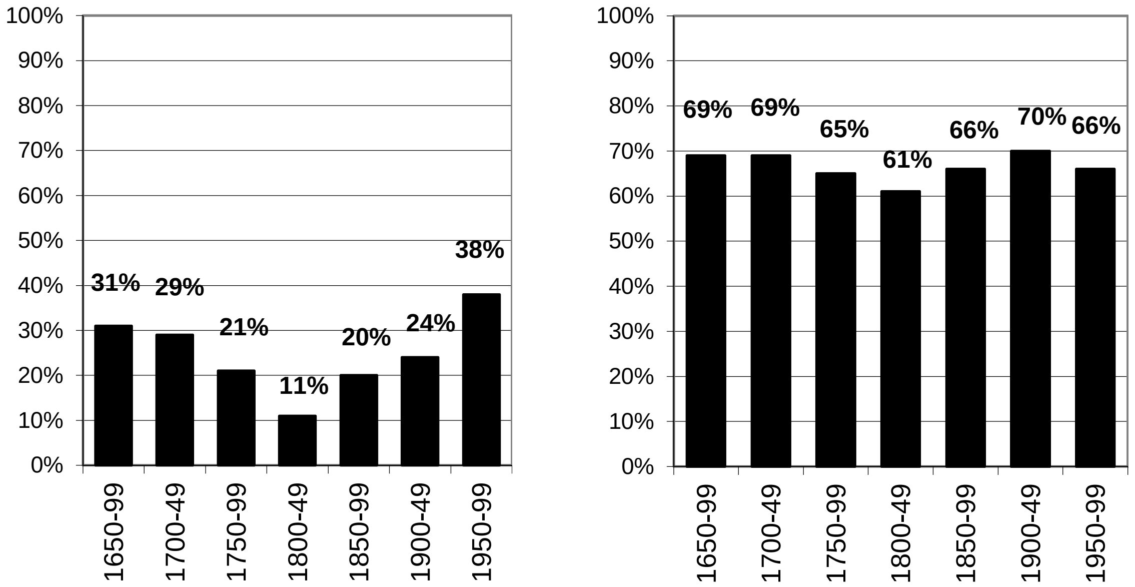

Figure 6. Rates of the *s*-genitive (left) and the ditransitive dative (right), on *y*-axis, by Archer real time period (*x*-axis).

The dependent variables in the Wolk at al. (2013) study are genitive realization (*s*- versus *of*) and dative realization (ditransitive versus prepositional). Their genitive dataset comprises *N* = 3,924 tokens, and their dative dataset *N* = 3,094 tokens. Figure 6 plots *s*-genitive and ditransitive dative rates against real time (categorized into 50-year periods, following ARCHER's corpus design). The dative alternation (right plot) is, as far as variant rates are concerned, rather stable in the Late Modern English period: ditransitive dative rates range between 61% (1800-1849) and 70% (1900-1949). By contrast, the genitive alternation (left plot) exhibits a good deal of frequency fluctuation. *S*-genitive rates started out at 31% at the beginning of the Late Modern English period, and fell subsequently to only 11% in the first half of the 19th century (1800-1849). After 1850, though, s-genitive rates recovered; in fact, the *s*-genitive is more popular, with a market share of 38%, in the second half of the 20th century (1950-1999) than ever. So one of the issues Wolk et al. (2013) are exploring is the extent to which (in)stability of probabilistic constraints is to blame for the variant rate trajectories depicted in Figure 6.

To statistically explore this question, Wolk at al. (2013) add a layer of rich contextual annotation to the datasets. The constraints for which they annotate the datasets include

- **length of possessor/possessum and recipient/theme** (in orthographic characters): these measures seek to do justice to the principle of end-weight, according to which e.g. long possessors should attract the *of*-genitive and long recipients the prepositional dative
- **animacy of possessor and recipient/theme** (up to five categories; human, collective, locative, temporal, inanimate): according to the literature, animate possessor should favor the *s*-genitive, and animate recipients the ditransitive dative
- **definiteness of possessor and recipient/theme** (up to 4 categories: indefinite, definite, proper name, (definite) pronoun): this measure is related to information structure
- **final sibilancy in possessor** (genitive alternation only): final sibilants in the possessor disfavor the *s*-genitive, as we have seen in Section 3.2
- **semantic relation** (genitive alternation only): according to the literature, prototypical genitive relations (e.g. kinship) favor the *s*-genitive

Additionally, each genitive and dative observation was annotated for real time. Based on this annotation ,Wolk at al. (2013) calculate two regression models to predict genitive and dative choices throughout the Late Modern English period; Figures 7 and 8 summarize the main effects in these models.

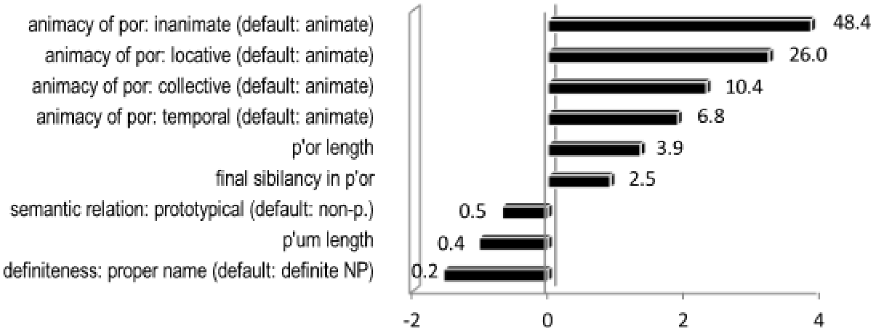

Figure 7. Genitive choice in Late Modern English: significant main effects in a minimal adequate mixed-effects logistic regression model. Bars plot regression coefficients (> 0: favoring, < 0: disfavoring); figures indicate odds ratios (> 1: favoring, < 1: disfavoring). Predicted odds are for the *of*-genitive. *N* = 3,824; correctly predicted: 92% (baseline: 76%), Somers Dxy = .93, κ = 8.4. Random effects: possessor head lemma, corpus file ID.

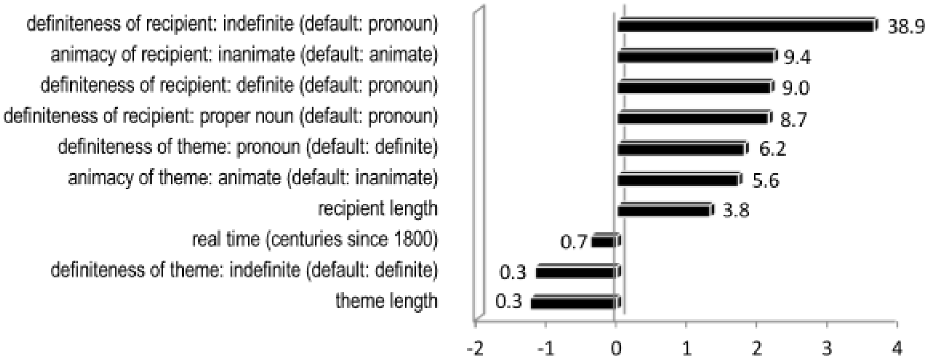

Figure 8. Dative choice in Late Modern English: significant main effects in a minimal adequate mixed-effects logistic regression model. Bars plot regression coefficients (> 0: favoring, < 0: disfavoring); figures indicate odds ratios (> 1: favoring, < 1: disfavoring). Predicted odds are for the prepositional dative. *N* = 3,093; correctly predicted: 94% (baseline: 66%); Somers Dxy = .97, κ = 7.2. Random effects: theme, verb lemma, register, corpus file ID.

In short, the main effects behave as expected, given the extensive literatures on the genitive (see, for example, Rosenbach 2002; Gries 2002; Hinrichs and Szmrecsanyi 2007) and dative (see, for example, Bresnan et al. 2007; De Cuypere and Verbeke 2013; Gries 2005) alternation. For instance, in the genitive model (Figure 7), an inanimate possessor increases the odds for the *of*-genitive by a factor of 48.8 – in other words animate possessors favor the *s*-genitive, as they should. Also, thanks to the principle of end-weight, longer possessums disfavor the *of*-genitive (and thus favor the s-genitive). As far as the main effects are concerned, Wolk et al.'s genitive model thus largely replicates the main effects in the late 20th century genitive model discussed in Section 3.2. As for the dative model (Figure

8), there are again few surprises: if the recipient is inanimate, the odds for the prepositional dative increase by a factor of 9.4, which is another way of saying that animate recipients attract the ditransitive dative, as reported in the literature. Conversely, long themes disfavor the prepositional dative and thus attract the ditransitive dative; this is again basically the theoretically expected end-weight effect.

The crucial question, however, is: are these effects probabilistically stable in real time? As we have seen in the previous two case studies, real time may have an effect in two ways: a syntactic variant is becoming more frequent over time regardless of context ("magical drift", see Section 3.1.), or real time interacts with the effect of language-internal predictors such that individual constraints become more or less important in the course of time (recall, e.g., the thematicity effect discussed in Section 3.2.). Recall also that only the latter scenario qualifies as a probabilistic change in the DPG approach. The analysis reported in Wolk at al. (2013) suggest that real time does not have a significant main effect in the genitive model; it does have a significant main effect in the dative model, however, indicating that the prepositional dative is becoming overall less popular in the course of the Late Modern English period. Importantly, though,, Wolk at al. (2013) also report interaction effects between language-internal predictors and real time with regard to both the genitive and the dative alternation. In the genitive model, these interactions are quite robust statistically; in the dative alternations, the interactions are a bit less robust but nonetheless clearly significant. In the DPG context of the present paper, then, we thus conclude that both the genitive alternation and the dative alternation have been subject to probabilistic change during the Late Modern English period.

Of the set of real-time interactions reported in Wolk at al. (2013), we take the liberty to focus here on a theoretically interesting cross-constructional parallelism: in both the genitive and the dative model, the effect that (some) animacy categories have on syntactic choices interacts significantly with real time. More specifically, the animacy constraint turns out to be subject to diachronic weakening: in the genitive model, starting in the middle of 19th century the *s*-genitive became less strongly disfavored with collective, locative, and temporal possessors. Wolk et al.'s dative model similarly suggests that inanimate recipients are coded significantly more often with the double object dative in the twentieth century than in earlier periods. In plain English, the *s*-genitive and the ditransitive dative construction have both come to be less "choosy" with regard to the animacy of the possessor/recipient, and this change happened at around the same time. Make no mistake: Late Late Modern English writers still preferably use both the *s*-genitive and the ditransitive dative with animate possessors/recipients, but this preference is probabilistically less strong now than it was at the outset of the Late Modern English period.

So in summary, we conclude that both dative and genitive grammars have evolved probabilistically in the course of the Late Modern English period – genitive grammars more so, dative grammars less so, but probabilistic change is implicated in both cases. We specifically discussed how in both alternations, the effect that more or less animate possessors or recipients have on syntactic choices has become weaker in the past 350 years. An investigation of the reasons for this change is beyond the scope of the present study; the reader is referred to the detailed discussion in Wolk at al. (2013) and Szmrecsanyi et al. (to appear). The important message is that language users' probabilistic knowledge of genitive and dative grammars has been subject to change.

4. Discussion and conclusion

Our starting point in this paper was converging experimental and observational evidence (see, for example, Bresnan 2007; Bresnan and Ford 2010) that language users have richer (i.e. probabilistic) knowledge of grammar than data from categorical grammaticality judgments would lead one to believe. This being so, Diachronic Probabilistic Grammar (DPG) endeavors to extend this insight backwards, and to infer linguistic knowledge of past language users from historical usage data (in other words, from historical corpora). Methodologically, we find in Probabilistic Grammar research a focus on variation, and specifically on the constraints that language users are demonstrably sensitive to when choosing syntactic variants. Hence, it stands to reason that DPG likewise centers on such constraints – more specifically, on the extent to which language users' sensitivity to variational constraints varies in real time. Thus, DPG's primary diagnostic of probabilistic change is the existence of significant interaction effects between real-time and language-internal predictors in regression analysis.

To scout out the diagnostic potential of DPG, we took a good look at three case studies featuring three well-known syntactic alternations in the grammar of English, and sought to establish if and to what extent these are subject to probabilistic change. We found that

- the alternation between non-finite and finite complementation (*I don't regret helping her start out* versus *I don't regret that I helped her start out*) is probabilistically fairly stable in the Late Modern English period, notwithstanding a non-alternation-specific, across-the-board drift towards increased usage of non-finite verb forms;
- the dative alternation (*I wrote M. an earnest loving note* versus *I wrote an earnest loving note to M.*) exhibits some probabilistic change in the Late Modern English period, especially with regard to the effect that recipient animacy has on syntactic choices;
- the genitive alternation (*the Seneschal's brother* versus *the brother of the Seneschal*) has been subject to substantial probabilistic change, in the Late Modern English period as well as more particularly in the late 20th century. This change has affected the animacy constraint, the thematicity constraint, the final sibilancy constraint, and the possessor length constraint.

At this point, a word on the relationship between frequency shifts and probabilistic change may be helpful. The fact of the matter is that frequency shifts may or may not be due to probabilistic change; conversely, probabilistic change may or may not involve frequency shifts. Accordingly, in regression analysis real time sometimes has a significant main effect (a phenomenon that we have dubbed "magical drift"), which crucially does not come within the remit of probabilistic change. The alternation between non-finite and finite complementation is a good example of a frequency shift without probabilistic grammar change: we have seen that the share of non-finite complementation has increased from about 50% in the earliest period we studied to about 70%, but we did not observe major probabilistic changes. Instead, it seems that the English language is simply drifting towards non-finite verb forms, regardless of context and probabilistic constraints. That said, recall that the dative alternation in the Late Modern English is frequency-wise fairly stable, yet subject to some probabilistic change. Of the variation phenomena we have explored in this paper, only the genitive alternation combines robust frequency shifts with substantial probabilistic change (see Szmrecsanyi 2013 for more discussion). The upshot is that probabilistic change and frequency shifts are two different beasts: frequency shifts may be due to any number of circumstances, and probabilistic change is just one of them.

Drawing on the Diachronic Probabilistic Grammar (DPG) framework this paper has ultimately sought to move diachronic corpus analysis beyond mere frequency analysis, exploring instead more or less subtle changes in the probabilistic conditioning of grammatical variation in the course of time. As Wolk et al. (2013, 414) succinctly put it, because we know that linguistic knowledge is to some extent probabilistic, in this endeavor "[h]istorical data is just another piece of evidence for how the mind works, another window to the mind".

References

Adger, David, and Graeme Trousdale. 2007. “Variation in English Syntax: Theoretical Implications.” *English Language and Linguistics* 11: 261–278.

Altenberg, Bengt. 1982. *The Genitive V. the Of-Construction. A Study of Syntactic Variation in 17th Century English*. Malmö: CWK Gleerup.

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. 2000. “Heaviness Vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering.” *Language* 76 (1): 28–55.

Behaghel, Otto. 1909. “Beziehungen Zwischen Umfang Und Reihenfolge von Satzgliedern.” *Indogermanische Forschungen* 25 (110-142).

Bod, Rens, Jennifer Hay, and Stefanie Jannedy, ed. 2003. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Bresnan, Joan. 1970. “On Complementizers: Toward a Syntactic Theory of Complement Types.” *Foundations of Language* 6 (3): 297–321.

———. 2007. “Is Syntactic Knowledge Probabilistic? Experiments with the English Dative Alternation.” In *Roots: Linguistics in Search of Its Evidential Base*, edited by Sam Featherston and Wolfgang Sternefeld, 75–96. Berlin, New York: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald R. Baayen. 2007. “Predicting the Dative Alternation.” In *Cognitive Foundations of Interpretation*, edited by Gerlof Boume, Irene Kraemer, and Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford. 2010. “Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English.” *Language* 86 (1): 186â€“213.

Bresnan, Joan, and Jennifer Hay. 2008. “Gradient Grammar: An Effect of Animacy on the Syntax of Give in New Zealand and American English.” *Lingua* 118 (2): 245â€“259.

Cuyckens, Hubert, Frauke D’hoedt, and Benedikt Szmrecsanyi. to appear. “Variability in Verb Complementation in Late Modern English: Finite Vs. Non-finite Patterns.” In *Late Modern English Syntax*, edited by Marianne Hundt. Cambridge: Cambridge University Press.

De Cuypere, Ludovic, and Saartje Verbeke. 2013. “Dative Alternation in Indian English: A Corpus-based Analysis: Dative Alternation in Indian English: A Corpus-based Study.” *World Englishes* 32 (2) (June): 169–184. doi:10.1111/weng.12017.

Denison, David. 1998. “Syntax.” In *The Cambridge History of the English Language*, edited by Suzanne Romaine, vol IV: 1776-1997:92–329. Cambridge: Cambridge University Press.

Ford, Marilyn, and Joan Bresnan. 2013. “Studying Syntactic Variation Using Convergent Evidence from Psycholinguistics and Usage.” In *Research Methods in Language Variation and Change*, edited by Manfred Krug and Julia Schlüter. Cambridge: Cambridge University Press.

Gahl, Susanne, and Susan Marie Garnsey. 2006. “Knowledge of Grammar Includes Knowledge of Syntactic Probabilities.” *Language* 82 (2): 405–410.

Gahl, Susanne, and Alan C.L. Yu. 2006. *Special Theme Issue: Exemplar-based Models in Linguistics*. The Linguistic Review. Mouton de Gruyter.

Givón, Talmy. 1980. “The Binding Hierarchy and the Typology of Complements.” *Studies in Language Groningen* 4 (3): 333–377.

Gries, Stefan Th. 2002. “Evidence in Linguistics: Three Approaches to Genitives in English.” In *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics*, edited by Ruth M. Brend, William J. Sullivan, and Arle R. Lommel, 17–31. Fullerton, CA: LACUS.

———. 2005. “Syntactic Priming: A Corpus-based Approach.” *Journal of Psycholinguistic Research* 34 (4): 365–399.

Gries, Stefan Th., and Martin Hilpert. 2010. “Modeling Diachronic Change in the Third Person Singular: a Multifactorial, Verb- and Author-specific Exploratory Approach.” *English Language and Linguistics* 14 (03): 293–320. doi:10.1017/S1360674310000092.

Hinrichs, Lars, Nicholas Smith, and Birgit Waibel. 2010. “Manual of Information for the Part-of-speech-tagged, Post-edited ‘Brown’ Corpora.” *ICAME Journal* 34: 189–231.

Hinrichs, Lars, and 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 (3): 437–474.

Huber, Magnus, Patrick Maiwald, Magnus Nissel, and Bianca Widlitzki. 2012. “The Old Bailey Corpus. Spoken English in the 18th and 19th Centuries.” *Www.uni-giessen.de/oldbaileycorpus*.

Hundt, Marianne, and Benedikt Szmrecsanyi. 2012. “Animacy in Early New Zealand English.” *English World-Wide* 33 (3): 241–263. doi:10.1075/eww.33.3.01hun.

Jäger, Gerhard, and Anette Rosenbach. 2008. “Priming and Unidirectional Language Change.” *Theoretical Linguistics* 34 (2): 85–113.

Labov, William. 1982. “Building on Empirical Foundations.” In *Perspectives on Historical Linguistics*, edited by Winfred Lehmann and Yakov Malkiel, 17â€“92. Amsterdam, Philadelphia: Benjamins.

Osselton, Noel. 1988. “Thematic Genitives.” In *An Historic Tongue: Studies in English Linguistics in Memory of Barbara Strang*, edited by Graham Nixon and John Honey. London: Routledge.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London, New York: Longman.

Rosenbach, Anette. 2002. *Genitive Variation in English: Conceptual Factors in Synchronic and Diachronic Studies*. Berlin, New York: Mouton de Gruyter.

———. 2005. “Animacy Versus Weight as Determinants of Grammatical Variation in English.” *Language* 81 (3): 613–644.

Szmrecsanyi, Benedikt. 2013. “The Great Regression: Genitive Variability in Late Modern English News Texts.” In *Morphosyntactic Categories and the Expression of Possession*, edited by Kersti Börjars, David Denison, and Alan Scott, 59–88. Amsterdam, Philadelphia: Benjamins.

Szmrecsanyi, Benedikt, Anette Rosenbach, Joan Bresnan, and Christoph Wolk. to appear. “Culturally Conditioned Language Change? Genitive Constructions in Late Modern English.” In *Late Modern English Syntax*, edited by Marianne Hundt. Cambridge: Cambridge University Press.

Tagliamonte, Sali. 2001. “Comparative Sociolinguistics.” In *Handbook of Language Variation and Change*, edited by Jack Chambers, Peter Trudgill, and Natalie Schilling-Estes, 729–763. Malden and Oxford: Blackwell.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. “Dative and Genitive Variability in Late Modern English: Exploring Cross-constructional Variation and Change.” *Diachronica* 3 (30): 382–419.

Yánez-Bouza, Nuria. 2011. “ARCHER Past and Present (1990â€“2010).” *ICAME Journal* 35: 205–236.

Zwicky, Arnold M. 1987. “Suppressing the Zs.” *Journal of Linguistics* 23: 133–148.