

Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora

LARS HINRICHS
Stanford University

and

BENEDIKT SZMRECSANYI
University of Freiburg

(Received 11 September 2006; revised 2 January 2007)

This study of present-day English genitive variation is based on all interchangeable instances of *s*- and *of*-genitives from the ‘Reportage’ and ‘Editorial’ categories of the ‘Brown family’ of corpora. Variation is studied by tapping into a number of independent variables, such as precedence of either construction in the text, length of the possessor and possessum phrases, phonological constraints, discourse flow, and animacy of the possessor. In addition to distributional analyses, we use logistic regression to investigate the probabilistic factor weights of these variables, thus tracking language change in progress as evidenced in the language of the press. This method, married to our large database, yields the most detailed perspective to date on frequently discussed issues, such as the relative importance of possessor animacy and end-weight in genitive choice (cf. most recently Rosenbach 2005), or on the exact factorial dynamics responsible for the ongoing spread of the *s*-genitive.

1 Introduction¹

1.1 *Variation between ‘s and of in genitival constructions*

Ever since its major phase of contact with French following the Norman Conquest,² the grammar of Standard English (StE) has had two competing ways of expressing a possessive relation between noun phrases: the inflected *s*-genitive and the analytical

¹ The authors’ names are given in alphabetical order, without any implication of priority. Funding from Deutsche Forschungsgemeinschaft, Bonn (grants no. MA 1652/3-1 and MA 1652/3-2), which made this project possible, is gratefully acknowledged. We thank Michael Percillier and Ulf Gerdemann for their substantial help with the coding. The audience of our talk at the ICAME 27 conference in Helsinki in May 2006, especially David Denison, provided very helpful advice on an earlier version of this article. We also benefited greatly and continuously from discussions with Christian Mair.

² Mustanoja (1960: 74) states, and Fischer (1992: 226) reaffirms, that the emergence of the periphrastic genitive using *of* is a native development that began in late Old English, similar to the development of the Latin preposition *de* into a genitive equivalent as used in Romance languages today. However, the spread of the *of*-genitive to its status as the clearly dominant form in Middle English was probably helped by contact with French. Mustanoja presents data (quoted from Thomas 1931 – Rosenbach 2002 quotes the same figures) suggesting that the *of*-genitive spread from very limited usage in learned writing (about 0.5 percent of all genitives) in the ninth and tenth centuries to about 85 percent of all tokens in the fourteenth century (1960: 75).

of-genitive.³ While the *s*-genitive can be considered a historical remnant of the Old English system of nominal cases,⁴ the postmodification of noun phrases with an *of*-prepositional phrase in the same function spread during Middle English, a change which was presumably supported by contact with the French model of the *de*-genitive.

Variation between the synthetic and the analytical construction is not free, i.e. the choice between the construction *President Kennedy's courageous actions* <Brown A06> and *the courageous actions of President Kennedy* is not entirely contingent. The constraints that govern speakers' and writers' choices between the two options have been addressed in a sizable body of research, and the contexts in which *s*- and *of*-genitive are interchangeable have received considerably more attention than those in which they are not (on which see e.g. Stefanowitsch 2003 for a construction-grammar perspective).

The best-known of these constraints on variation is the animacy constraint, according to which the *s*-genitive is preferred if the possessor is animate. While prescriptivist grammars of English tend to recommend the use of the *s*-genitive with animate, personal possessors,⁵ our corpora also provide numerous examples of *inanimate* possessors taking the *s*-genitive, cf. *the executive mansion's library* <Brown A33>, *the earth's atmosphere* <Brown A16>, *the building's precarious state* <Frown A26>. One question we will address is whether the *s*-genitive has actually been spreading to inanimate possessor head nouns (cf. Mair 2006a).

Further constraints that have been discussed by large numbers of writers relate to information status – suggesting that if the possessor noun phrase is in some way given, known to the reader, or more relevant to the text at hand than other nouns, the *s*-genitive will be preferred – and to the principle of end-weight, which states that language users will prefer the type of genitive in which the longer of the two noun phrases occurs second. Since the order of possessor and possessum are converse in the two types of genitive, this principle potentially has strong bearing on genitive variation (cf. figure 1).

There is no consensus in the literature on the relative importance of the various factors that seem to be influencing genitive choice. The debate over the status of animacy relative to the end-weight principle is a case in point: while Hawkins (1994) proposes that the role of animacy and its favoring effect on choice of the *s*-genitive is ultimately

³ Following Rosenbach (e.g. 2002, 2003, 2005, 2006) and others, we refer to both the *s*- and the *of*-form as 'genitive' on functional and semantic grounds. However, we are aware that some authors draw a terminological distinction between the *s*-genitive' and the *of*-construction' (Biber, Johansson, Leech, Conrad & Finegan 1999a, e.g. Kreyer 2003 and standard reference grammars, e.g. Quirk, Greenbaum, Leech & Svartvik 1985).

⁴ As Mair (2006a) points out, it has been suggested that the modern 's-marker should not really be regarded as a direct continuation of the genitive inflection of the Old English strong masculine class of nouns but as a clitic. Among the facts pointed to in support of such an analysis are group genitives or the absence of *s*-voicing (cf. the plural *wives* as opposed to the genitive *wife's*). For a critical review of the arguments in the debate see Allen (2003) and Rosenbach (2002).

⁵ Typically, usage guides give recommendations rather than clear rules on genitive choice, e.g.: 'prefer the possessive pattern *X's Y* in the following conditions: . . . when *X* describes a person rather than a thing' (Leech, Cruickshank & Ivanič 2001: 406); see also Burchfield (1996: 688). The lack of clarity in this area of language use may be the reason why other guides avoid the topic of choosing between 's and *of*; see e.g. Peters (2004), Swan (1995).

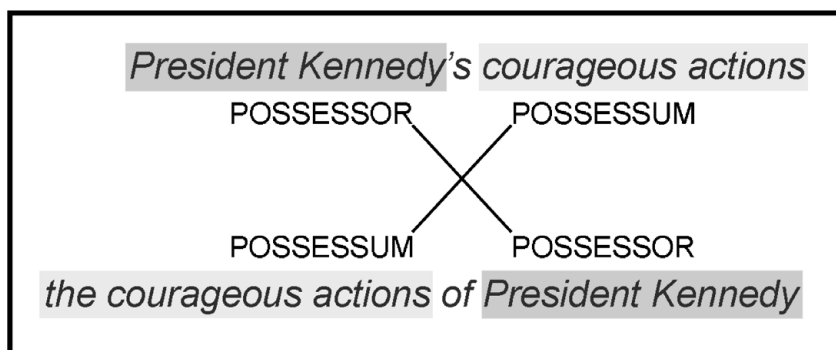


Figure 1. Converse positions of possessor and possessum in the *s*-genitive (above) and the *of*-genitive <Brown A06>

epiphenomenal to end-weight (because animate possessors tend to occur as proper nouns and are therefore usually shorter than inanimate common nouns), Rosenbach (2005) argues that animacy has an effect on genitive choice that is independent of end-weight.

Few, if any, previous studies of genitive variation have discussed the relative weights of any more than two different factors in quantitative terms.⁶ Mair (2006b) points to the methodological challenge that is posed by the detailed investigation of genitive variation:

the determinants of synchronic variation in genitive usage interact in complex ways, and ... it is very difficult to identify diachronic trends against a background of sometimes far greater synchronic variability. (Mair 2006b: 146)

Clearly, what is most needed in this debate, which has involved variationists and corpus linguists, is a comprehensive corpus-based study determining the contribution of factors, and their relative strength, for a broad range of constraints that purportedly influence genitive choice. That is what the present study offers, albeit limited to journalistic language.

Thus, we will seek to operationalize a wide array of factors and relate them to the findings reported in previous scholarship on genitive variation. The factors are grouped in four major sections: (i) semantic and pragmatic factors; (ii) phonological factors; (iii) factors related to processing and parsing; (iv) economy-related factors. Section 5 will present a factor-by-factor distributional analysis. Section 6 will model the joint probabilistic impacts of these factors on genitive choice, using logistic regression as a multivariate analysis method. Our approach affords insights into the structure of synchronic variation in British versus American English genitive marking;

⁶ Szmrecsanyi (2006: 87–107) presents a multivariate analysis of constraints on genitive choice in spoken English, albeit with a primary theoretical interest in persistence effects, not genitive variation. Leech, Francis & Xu (1994) offer a multivariate analysis of the effects of three factors on genitive choice: animacy, genre, and the semantic relation between possessor and possessum. Their study draws on parts of LOB for data.

diachronically, we will be able to thoroughly review the observations that various authors have made regarding the ongoing shift away from the *of*- and toward the *s*-genitive. Linguists have been noting this shift since the early twentieth century (to name some of the earlier sources: Barber 1964: 132–3; Jespersen 1909–49: VII, 327–8; Potter 1969: 105); this literature suggests that the development originated in journalistic language and spread to other text types from there.⁷ Our data support the view that there is a shift among the two constructions for journalistic language in the period from 1961 to 1991/2 (cf. section 4).

There is no consensus whether the shift from *of* to *'s* is due to changes in the animacy constraint: some authors attribute it to a spread of the form to inanimate possessor noun phrases (tentatively also Denison 1998, e.g. Jespersen 1909–49: VII, 327–8),⁸ while Mair (2006a, 2006b) claims that the animacy constraint is currently being loosened for collective nouns, not inanimates, and that furthermore, the more significant causes of the spread of the *s*-form lie in the area of discourse practices,⁹ not the underlying constraint grammar (2006b: 147).

The overarching aim of this article, then, is to pinpoint the constraints that are responsible for this shift, i.e. those that have lost or gained explanatory power. Until now, the question of the causes of this shift – as Mair points out, a remarkable ‘partial reversal of a general drift towards analyticity in English grammar’ (2006b: 146) – remains wide open.

1.2 Background

Previous work on corpora of the English language, including considerable amounts of research based on the Brown family of corpora, have studied diachronic and synchronic variation. Certain theoretical concepts have thus been arrived at in ‘bottom-up’ approaches which have proven useful in the description of both journalistic prose and other genres. This section briefly presents some of those central ideas, as the interpretation of our results will benefit from their application.

1.2.1 Journalistic prose as an ‘agile’ genre: responses to the demands of ‘popularization’ and ‘economy’

Douglas Biber and Edward Finegan have shown in their studies of register variation in the historical ARCHER corpus (Biber & Finegan, 1989) that

⁷ See also Altenberg (1982: 15) who observes ‘signs of a renaissance in the use of non-personal genitives . . . especially in . . . headlines and journalese’.

⁸ However, our data put us in no position to discuss whether newspaper language is actually the *origin* of the shift. This claim has not yet been tested corpus-linguistically. Interestingly, one publication that supported it in its first edition, Fowler’s *Dictionary of English Usage* (1926), turned away from the claim in its more recent editions: ‘The reason for the shift in this direction lies deeply buried in a long-drawn-out historical process. Newspaper headlines, *pace* Fowler, have had little or nothing to do with it’ (Burchfield, 1996: 688–9).

⁹ The factors which Mair sees as playing the biggest role in the shift are *horror aequi* (cf. section 5.3.3 of this article) and economy, i.e. the use of the *s*-genitive as the more compact device for information packaging (cf. section 5.4).

[w]ritten prose registers in the seventeenth century were already quite different from conversational registers, and those registers evolved to become even more distinct from speech over the course of the eighteenth century. (Biber 2003: 169)

However, beginning in the nineteenth century and as a consequence of increasing literacy and the ascendance to power of an educated bourgeoisie in England, certain popular written genres (letters, fiction, essays) ‘reversed their direction of change and evolved to become more similar to spoken registers’. Most notably, these genres started displaying a dispreference for certain ‘stereotypically literate features, such as passive verbs, relative clause constructions and elaborated noun phrases’ – i.e. those forms which became more frequent in the academic genres. This dissimilation of text types still continues, as Biber & Finegan (2001) have shown; in fact it accelerated notably in the twentieth century.

In his study of (British) newspapers, Biber (2003) demonstrated that the writing of journalists is located at a very delicate genre-typological place, right between the more popular and the more literate genres that Biber and Finegan showed to be undergoing dissimilation. Modernity has caused an ‘informational explosion’ (Biber, 2003: 180), and the amounts of knowledge that have to be transmitted by informational texts such as newspapers still keep growing every day. Thus, in Biber’s terms, the pressure of economy increases continuously. This development and its reflection in journalistic language started to accelerate during the ‘last fifty to one hundred years’ (2003: 180), Biber argues in his synchronic comparison of newspaper texts with other genres (conversation, fiction, and academic prose). Using the Longman Corpus of Spoken and Written English, he finds that those textual features which help convey information in a compact way are most dominant in news language.

Biber’s study complements an earlier publication by Hundt & Mair (1999). While Biber focused on demonstrating how press language is sensitive to the demands of economy, Hundt & Mair showed in a similar typology of genres that newspapers, compared to the other text types they considered, are most likely to exhibit innovative forms of language use, including ‘changes from “below”’ (Hundt & Mair 1999: 235). Their study of the 1961–91/2 time span – using the (non-POS-tagged version of the) same set of corpora as the present article – demonstrates a strong uptake for colloquial variables by newspaper language. According to Hundt & Mair, press texts are therefore the most ‘agile’ genre of all. In their interpretation, which Biber seconds, this is a response of journalistic prose to the pressures of the market, designed to win wider audiences through the use of a more ‘involved’ writing style (cf. Biber 1988).

Read together, Biber (2003) and Hundt & Mair (1999) demonstrate that the linguistic responses to the demands of both popularization and economy are defining developments which need to be considered in a study of English newspaper language. Also, newspaper prose seems to be the most promising genre to analyze in any study of language change in progress, given its openness to innovation. However, as Hundt & Mair point out, one should be careful not to generalize the findings from a newspaper corpus to other genres, considering this special place of press language in the spectrum of genres (Hundt & Mair 1999: 236).

1.2.2 *'Colloquialization' and 'Americanization'*

Outside of the specific dynamics of press language, the Brown family of corpora has repeatedly shown patterns of variation that have been described as colloquialization and Americanization. The first of these refers to the phenomenon of changes through which written language becomes more similar to spoken language (see Hundt & Mair 1999: 225–6 for a detailed definition). This tendency of a partial rapprochement between spoken and written norms in late modernity had been observed earlier by students of the sociocultural context and described as a 'democratization' of the written norm (Fairclough 1992: 221). Comparisons between corpora from different points in time such as Brown vs. Frown or LOB vs. F-LOB have substantiated the existence of such a drift by noting growing frequencies of 'involved' features such as semimodals and the progressive aspect (Krug 2000 – cf. also above for the features treated in Hundt & Mair 1999), or shrinking frequencies of 'informational' features such as the passive voice (Leech & Smith 2006).

'Americanization' refers to a frequently corresponding phenomenon that has been found to be at work for many of these features: alternatively called the 'follow-my-leader' pattern of British English (BrE), it typically describes processes of colloquialization which are led by American English (AmE), i.e. for which (AmE) shows higher frequencies of a colloquial variant in the 1961 and 1991/2 corpora than BrE (or smaller frequencies of a formal variant), and/or displays a greater rate of change toward a more colloquial variant than BrE. Such a leading role of AmE is demonstrated by the decline in core modals and a corresponding increase in semimodals reported by Leech & Smith (2006: 189).

1.3 *Research objectives*

We will interpret our findings in the light of the concepts laid out in the previous section. As our central task we will have to discuss whether the variation we observe in the area of the English genitive is indeed an instance of colloquialization, or whether it is inadequately placed in this framework. Leech & Smith (2006) claim that the increase of the *s*-genitive in the Brown family of corpora, along with a 'roughly commensurate' loss of *of*-genitives in a subsample of the corpora, 'fits into the mould of colloquialization' (197). This argument presupposes a clear functional split between '*s*' as the more informal variant and *of* as the more formal variant, a view which, they argue, is permissible based on the fact that the strong increase in usage of the *of*-genitive in general English was preceded by increased use in the language of educated writers in ME.¹⁰ However, given the considerable potential of '*s*' as the more condensed,

¹⁰ Fischer & van der Wurff (2006: 118) point out that the *of*-genitive is probably native to English, as it is to other Germanic languages. The effect of contact with French was a strong increase in frequency of usage. The most forceful argument in favor of '*s*' being the more informal choice is probably made by Altenberg (1982), who points to the 'strong OF preference in . . . formal contexts' in his corpus of seventeenth-century written language which goes back to usages introduced by such religious writers as Wycliffe and Purvey (255–6; e.g. such phrases as *the power of God, the Name of the Lord, the body of Jesus*).

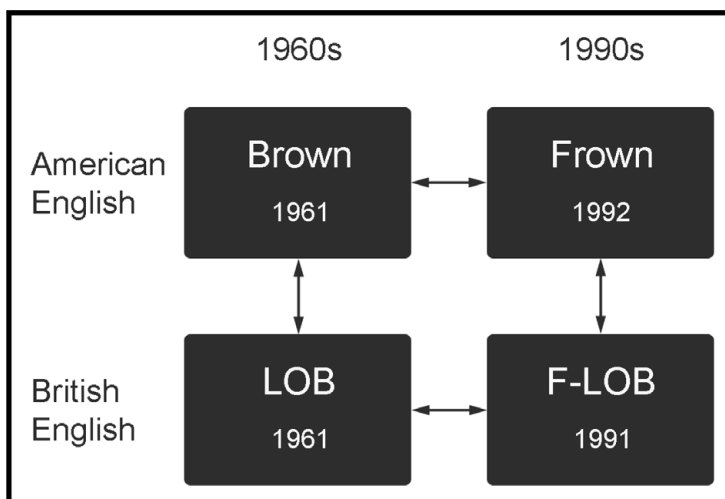


Figure 2. The Brown quartet of matching corpora of written Standard English

abstract, ‘informationally’ oriented variant, a multivariate analysis such as ours will help to distinguish between those aspects of genitive variation that can actually be ascribed to colloquialization, and those which might be better explained as, for example, economization strategies (see our discussion of this aspect in section 7.3 below).

In short, our research objectives in this article are:

- (i) to determine the hierarchy of factors that influence genitive choice in journalistic language, based on the analysis of all four corpora of StE,
- (ii) to explore, and account for, differences in genitive choice between BrE and AmE,
- (iii) to model the ongoing shift from *of-* to *s-*genitives in press language in terms of changing weights associated with noncategorical constraints in a probabilistic grammar framework (cf., for instance, Bresnan, Cueni, Nikitina & Baayen, forthcoming; Manning, 2003),

On the methodological plane, it follows naturally from the above that we will adopt a variationist approach to genitive variation, in the spirit of, for example Labov (1969) and Weiner & Labov (1983). In this connection, we will seek to demonstrate the value of part-of-speech-tagged (POS-tagged) corpora in combination with multivariate variationist methodology.

2 The data

Our choice of data is press material (sections A and B) in the Brown family of corpora, a set of four corpora of written StE documenting two varieties of English at two different points in time: British English and American English in the 1960s and 1990s (see figure 2). All corpora were compiled according to the design of the first corpus, Brown,

which comprises fifteen genre categories that contain a total of 500 text samples of 2,000 words each, amounting to a total of one million words per corpus (see Appendix).¹¹

Following the compilation of Brown and LOB at, respectively, Brown University and the Universities of Lancaster, Oslo, and Bergen, the compilation and the POS-tagging of the two newer corpora, F-LOB and Frown, was performed in cooperation by the Universities of Lancaster and Freiburg in teams headed by Geoffrey Leech and Christian Mair, respectively. After having automatically assigned a grammatical tag to each word in the corpora using the C8 tagging suite, tags in the two more recent corpora were manually postedited at Freiburg in order to minimize the risk of erroneous tagging. Thus, the four corpora are now available with grammatical markup of rather high quality.¹²

There is already a large and growing body of literature presenting research based on these data. Some central publications containing detailed information on the processes of compilation and the markup are Francis & Kučera (1982) on Brown, the first of the four corpora to be completed, documenting written American English of the early 1960s; Johansson & Hofland (1989) on LOB, the follow-up corpus designed to match Brown for British English; Sand & Siemund (1992) as well as Hundt, Sand & Siemund (1998) on F-LOB, the 1990s update of LOB; and Hundt, Sand & Siemund (1999) on Frown, the 1990s update of Brown. Mair et al. (2002) presented a first report on POS-frequency shifts from LOB to F-LOB.

3 The linguistic variable

Following standard practice in the variationist literature, this section will circumscribe the variable context – as Tagliamonte & Smith succinctly put it, an ‘accurate delimitation of the variable context is critical, as inclusion or exclusion of certain contexts may skew the data’ (2002: 262). To rule out such skewings, all instances of interchangeable *s-* and *of-*genitives were extracted from the subcorpora, i.e. each instance of an *s-* or *of-*genitive was classified according to whether the alternative construction could have been used in its place (see below for a discussion of the criteria employed in the selection process). Table 1 gives the number of all *s-*genitives and *of-*constructions that were considered for each of the four subcorpora, as well as the

¹¹ Currently, the set of corpora is being expanded at Lancaster University by the compilation of two more corpora representing British English around the years 1900 and 1960.

¹² LOB has been available in a part-of-speech-tagged version that was already postedited by the original team that produced it (cf. Johansson & Hofland 1989). Tags were initially assigned using the ‘LOB tagging suite’ (Johansson, Atwell, Garside & Leech, 1986), manually postedited, and then automatically mapped onto the current version of the C8 tagset by Nick Smith; it can therefore be considered to have the same level of error-freeness as F-LOB and Frown. Brown has been automatically marked up using C8 and has not yet been postedited. This makes for minor error margins (the automatic tagger output can be considered to be about 98 percent correct). For the present study all tokens that entered analysis were hand-selected; only a minimal number of tokens will have been overlooked due to erroneous tagging in Brown.

Table 1. *Raw frequencies of genitival <'s>, <s'> and of versus the number of tokens selected as 'interchangeable genitives'*

	Brown A/B*	LOB A/B	Frown A/B	F-LOB A/B	TOTAL
Total number of <i>s</i> -genitives	995	947	1,377	1,300	4,619
Number of tokens coded as interchangeable	80% (797)	80% (756)	82% (1134)	69% (891)	77% (3578)
Occurrences of the preposition <i>of</i>	4,582	4,363	3,683	3,796	16,424
<i>of</i> -tokens following a noun and not preceding a verb	2,264	2,282	1,860	1,944	8,350
Number of tokens coded as interchangeable	31% (1407)	29% (1263)	27% (998)	28% (1054)	29% (4722)
Total interchangeable: $N = 8,300$					

*From each corpus, the two largest of three press text categories were selected: A 'Reportage' (44 text samples per corpus) and B 'Editorial' (27 samples).

number of tokens selected and coded for analysis. Altogether, the study is based on a data set of $N = 8,300$ interchangeable genitives.

The preparation of the datasets relied strongly on the available POS-tagging in the corpora. Since the C8 tagger assigns a discrete tag to *s*-genitives,¹³ these were easily retrieved. By comparison, the selection process with untagged data would have been extremely laborious as the *s*-genitives would need to be separated from plural forms of nouns, nouns ending in cliticized forms of *be*, etc.¹⁴

In a final step, the genitive tokens thus preselected were manually coded for interchangeability. We retained only those instances of the inflected *s*-genitive which could plausibly have been expressed as an *of*-genitive by applying a simple conversion rule, without adding or deleting any of the lexemes in the possessor or possessum phrase (except for the optional addition of a determiner to the possessum). Similarly, only those *of*-genitive tokens were retained which could have been expressed using an *s*-genitive construction instead with neither of the noun phrases modified, except for the necessary deletion of any determiner in the possessum phrase. Crucially, the alternative

¹³ The tag system marks all instances of regular genitival '*s*' (*dog's*) as well as 'bare genitives' (*dogs*) (Huddleston & Pullum 2002: 1595–6; Kaye 2004).

¹⁴ Extracting all instances of genitival *of* was a more complicated process. First, the very large number of *of*-tokens in the data was automatically scanned for instances of *of* which were (a) preceded by a word tagged as noun and (b) *not* followed by a word tagged as verb. From these remaining instances, several frequent, nongenitival constructions were then eliminated. Thus, 29 percent of all occurrences of *of* were tagged as parts of interchangeable genitive tokens.

construction would have to leave the meaning of the actual choice unchanged; thus, *the city of Atlanta* was not considered an interchangeable genitive because the alternative, *Atlanta's city*, has a different meaning.

Our notion of 'interchangeability' does not imply that the *s*-genitive and the *of*-genitive would have been equally 'felicitous' in the context at hand (some authors use this notion in discussions of genitive variation, e.g. Dixon 2005). As many English-language writers would agree, *Mrs. Eustis Reily's olive-green street length silk taffeta dress* <Brown A18> is a more felicitous wording than the one using the alternative genitive construction, *the olive-green street length silk taffeta dress of Mrs. Eustis Reily*. Nonetheless, we considered this *s*-genitive token to be interchangeable because a conversion to the alternative construction would not have significantly altered its meaning. The aim of our analysis is indeed to explain what makes one genitive construction more felicitous than the competing one.

A negative list of noninterchangeable types and cases guided the coders' judgments of interchangeability. While *s*-genitives proved to be interchangeable in the great majority of cases, the following were excluded from the analysis:¹⁵

- (i) any construction in which a noun marked with a genitive *s* is not followed by an explicit possessum phrase, since any transformation to a postmodified noun phrase would require the addition of lexical items, or would yield a phrase introduced by a different preposition than *of*. These are the noninterchangeable contexts described by Kreyer (2003: 170): 'independent genitives'¹⁶ (*Her memory is like an elephant's*) (Quirk et al. 1985: 329); 'local genitives' (*Let's have dinner at Tiffany's*) (329–39); and 'post-genitives'¹⁷ (*a friend of Jim's*) (330–1);
- (ii) any phrase that has been conventionalized with the *s*-genitive, so that the *of*-genitive is no longer a possible alternative (*Murphy's law*);
- (iii) 'descriptive genitives'¹⁸ (*men's suits, bird's nest*), which frequently form an idiomatic unit (Quirk et al. 1985: 327–8) and are therefore excluded by virtue of criterion (ii) above,¹⁹ and/or would not take *of* if transformed into a noun phrase with a prepositional postmodifier;
- (iv) any *s*-genitive construction whose possessum noun phrase is premodified by *own* (*the president's own agenda*);
- (v) titles of books, films, works of art, etc. that are premodified with a genitive possessor phrase denoting their creator, since a transformation would require a *by*-phrase rather than an *of*-phrase (*John Steinbeck's Of Mice and Men*).

¹⁵ Compare Kreyer's (2003: 170) and Rosenbach's (2006: 622–3) similar sets of criteria for interchangeability of genitives.

¹⁶ Alternatively, 'elliptic genitives' (Biber, Leech & Johansson 1999b: 296–7) or Huddleston & Pullum's type III: 'fused subject-determiner-head' (2002: 468).

¹⁷ Alternatively, 'double genitives' (Biber et al. 1999b: 299) or Huddleston & Pullum's type IV: 'oblique genitives' (2002: 468–9).

¹⁸ Alternatively, 'classifying genitives' (Biber et al. 1999b: 294–5) or Huddleston & Pullum's type VI: 'attributive genitives' (2002: 469–70).

¹⁹ See Rosenbach (2006) for a recent discussion of gradience between this type of genitive and compound-like noun + noun sequences.

The types of *of*-genitives that were excluded from the analysis included the following:

- (i) *of*-genitive constructions with a possessum that could not possibly be read as definite, since the *s*-genitive always expresses the possessum as definite. Thus, cases such as the following were excluded: *a major strategy of his administration, some members of his cabinet*;
- (ii) most of those *of*-genitives containing a possessor noun phrase that shows postmodification, since the result of a conversion to the *s*-genitive would be a ‘group genitive’ (*she’s the second guy from the right’s sister*) (Quirk et al. 1985: 328).²⁰ Group genitives are proscribed in written StE, and certainly in edited writing such as newspaper writing. Therefore, the *s*-genitive could not be considered a possible competing choice in these instances. However, if the postmodifier was reasonably short and tightly integrated with the head noun to form a conventionalized unit (*the University of Arkansas, the Museum of Modern Art*), so that the affixation of ‘s to the postmodifier would not have broken good stylistic usage rules, coders were free to decide that the *of*-genitive was in fact interchangeable with an *s*-genitive;
- (iii) measures expressed as *of*-constructions, as in *a pound of flesh, fourteen days of rain*;
- (iv) as with *s*-genitives, any phrase that has been conventionalized and spread with an *of*-genitive (*the University of Mississippi, the President of the United States*).

Outside of these lists of negative cases, coders relied on their own judgment. The four coders who were involved in classifying all occurrences of *s*-genitives and *of*-genitives in the data as either interchangeable or not interchangeable all received coder training. To set a bound on error levels, to ensure replicability of the findings, and to enhance confidence in the coding scheme, the procedure laid out in Orwin (1994) was followed and Cohen’s κ , which measures intercoder reliability by establishing the proportion of the best possible improvement over chance, was computed. Prior to the coding of the main sample, a number of random genitive samples from the data were independently coded by two of the four coders for interchangeability of the genitives. After a number of trials on different samples and a series of subsequent refinements to the coding scheme, annotation of a set of $N = 66$ *s*-genitives and $N = 136$ *of*-genitives yielded (i) a simple agreement rate of 86 percent and a ‘good’ (cf. Orwin 1994: 152) Cohen’s κ value of 0.69 for *s*-genitives, and (ii) a simple agreement rate of 89 percent and an ‘excellent’ Cohen’s κ value of 0.78 for *of*-genitives. This means that our coding scheme is sufficiently dependable, and that intercoder reliability of our annotation is satisfactory.

Following the identification of the basic data set of genitive tokens, further manual and automatic coding was applied. The human coders established the boundaries of all tokens by adding distinct marks at the beginning of the possessor phrase and at the end of the possessum phrase for *s*-genitives, and vice versa for *of*-genitives. Also, each token was coded for animacy of its possessor (see section 5.1.1 on the animacy scale employed). Using Perl (Practical Extraction Resource Language) scripts, the tokens as well as possessor/possessum boundaries were automatically retrieved and annotated for a number of additional conditioning factors, which form the basis of our analysis.

²⁰ Or ‘phrasal genitives’ (Huddleston & Pullum 2002: 479–80).

Table 2. *Share of the s-genitive out of all interchangeable s- and of-genitives by corpus*

	<i>s</i> -genitive		Total
	%	<i>N</i>	<i>N</i>
Brown A/B	36.2	797	2,204
LOB A/B	37.4	756	2,019
Frown A/B	53.2	1,134	2,132
F-LOB A/B	45.8	891	1,945
Total	43.1	3,578	8,300

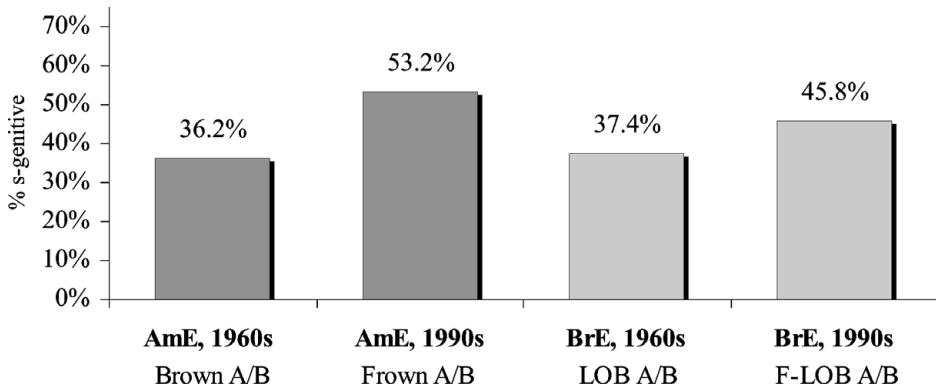


Figure 3. Share of the *s*-genitive of all interchangeable genitives by corpus

These values will be addressed in more detail in appropriate places below (see sections 5 and 6), which present the results of this study.

4 Overall distribution of genitives

We will first explore the overall distribution of *s*-genitives and *of*-genitives in the data (table 2 and figure 3). We can observe that since the 1960s, the relative frequency of the *s*-genitive has increased substantially in both BrE (37 percent to 46 percent) and – even more markedly – in AmE (36 percent to 53 percent); both of these increases are significant at the $p < .001$ level, according to a chi-square test of independence. Our least surprising finding should thus be that the *s*-genitive is indeed spreading, given the many claims to this effect in the literature (for instance, Dahl 1971: 141; Potter 1969: 105–6; Raab-Fischer 1995; Rosenbach, 2003: 394–5). In this context it should be pointed out that while the difference in the share of *s*-genitives between LOB and Brown is not statistically significant, the difference between F-LOB and Frown (and thus between recent written BrE and AmE) is ($p < 0.01$). Again, it has been noted before that the *s*-genitive is more frequent in AmE than in BrE (for instance, Rosenbach 2003: 394–5), but the clear pattern of divergence exhibited in our data is, we believe, striking.

Table 3. *Coding scheme for animacy*

	coding value	examples	tag in Zaenen et al.'s (2004) scheme
human	"1"	<i>girl, Jones</i>	HUMAN
animal	"2"	<i>dog, horse</i>	ANIMAL
collective	"3"	<i>the UN, party</i>	ORG
inanimate	"4"	<i>chair, morning</i>	PLACE, TIME, CONCRETE, NONCONC, MAC, VEH

5 Conditioning factors and distributional analysis

This section will engage in a detailed discussion of the conditioning factors considered in the present study (and the way they were coded). Along the way, we will report a series of univariate analyses of how these individual conditioning factors interact with genitive variation in the data. As has been explicated earlier, the factors considered fall into four groups: (i) semantic and pragmatic factors, (ii) phonological factors, (iii) factors related to processing and parsing, and (iv) economy-related factors.

5.1 *Semantic and pragmatic factors*

This factor group comprises factors that relate to the lexical class of the possessor (i.e. animacy), to the pragmatic status of the possessor in a given corpus text (i.e. thematicity of the possessor), and information status (or discourse flow) as a higher-level pragmatic factor.²¹

5.1.1 *Animacy*

The rich literature on genitive variation agrees that the lexical class of the possessor is the most crucial conditioning factor for predicting genitive choice. Hence, the more human and animate a possessor, or the more it conveys the idea of animate things and human activity, the more likely it is to take the *s*-genitive (cf. Altenberg 1982: 117–48; Biber et al. 1999b: 302–3; Dahl 1971: 140; Jucker 1993: 126–8; Kreyer 2003: 172; Rosenbach 2003, 2005; Taylor 1989: 668–9). Adopting Rosenbach's (2006: 105) animacy hierarchy (human > animal > collective > inanimate) and drawing on Zaenen et al.'s (2004) general coding scheme for animacy, we sought to operationalize the factor 'animacy' by manually coding each possessor NP in our database according to the four-way classification shown in table 3.

Two coders, both trained linguists, were involved in coding the database. To determine reliability of the coding decision, a random subset of the data ($N = 199$ possessor NPs) was classified independently by both coders. Cohen's κ was again

²¹ Owing to our variationist assumption of rough semantic interchangeability between the two genitive constructions, this study will not consider the semantic relation – possessive, subjective, objective, and so on (cf. Quirk et al. 1985: 321–2) – between the possessor and possessum phrase as a conditioning factor (notice here that according to Leech, Francis & Xu 1994, this factor is not exceedingly important anyway compared to other constraints, such as animacy).

Table 4. *Mean possessor animacy by corpus and genitive type*

	<i>s</i> -genitive		<i>of</i> -genitive	
	mean	std. dev.	mean	std. dev.
Brown A/B	2.22	1.30	3.15	1.15
LOB A/B	2.28	1.18	3.24	.95
Frown A/B	2.42	1.34	3.25	1.17
F-LOB A/B	2.31	1.16	3.29	1.01
Total	2.32	1.26	3.23	1.07

computed to evaluate intercoder reliability, yielding a simple agreement rate of ca. 86 percent and an ‘excellent’ (Orwin 1994: 152) κ value of ca. 0.79.

Table 4 crosstabulates mean animacy scores with corpus and genitive type. Overall, *s*-genitive possessors are clearly more animate than *of*-genitives possessors, a difference which an independent sample t-test shows to be highly significant ($p < .001$). More specifically, the typical *s*-genitive possessor is located almost one notch higher (at 2.32) on Rosenbach’s (2006) animacy hierarchy than the typical *of*-genitive possessor (3.23). Needless to say, this finding squares with the literature. Second, the standard deviation associated with animacy of *s*-genitive possessors (1.26) is higher than the one associated with animacy of *of*-genitive possessors (1.07), which is another way of saying that the *s*-genitive is more versatile, in terms of the lexical class of its possessor, than the *of*-genitive. What about longitudinal changes? A univariate analysis seems to suggest that the *s*-genitive in particular has come to be associated, over time, with more inanimate possessors: the mean value in our 1960s sample is 2.25 while it is 2.37 in our 1990s sample ($p < .005$). This is consonant with claims in the literature that *s*-genitives with inanimate possessors are ‘on the increase’ (Denison 1998: 119). Along these lines, note that *s*-genitive possessors in Frown are significantly ($p < .05$) less animate than possessors in F-LOB.

5.1.2 *Thematic genitives: text frequency of the possessor head*

Osselton (1988) has claimed that it is the general topic of a text which determines, among other things, which nouns in that text can take the *s*-genitive. Thus, according to Osselton, while *sound*, *soil*, and *fund* will not normally take the *s*-genitive, ‘in a book on phonetics, *sound* will get its genitive, in one on farming, *soil* will do so, and in a book on economics you can expect to find *a fund’s success*’ (Osselton 1988: 143). We aimed to operationalize Osselton’s notion of ‘thematic genitives’ by having a Perl script establish, for every individual possessor NP in our database, the text frequency of the possessor NP’s head noun in the respective corpus text, assuming that the more central thematically a given noun is in a given text, the more often it will occur in that text. Let us illustrate the procedure with the example in (1):

Table 5. *Mean text frequency of the possessor head noun by corpus and genitive type*

	<i>s</i> -genitive		<i>of</i> -genitive	
	mean	std. dev.	mean	std. dev.
Brown A/B	5.23	5.90	3.90	4.42
LOB A/B	5.27	5.24	4.41	5.59
Frown A/B	7.18	7.25	3.76	4.60
F-LOB A/B	5.55	4.90	3.31	5.32
Total	5.94	6.07	3.88	5.01

- (1) *The bill's supporters* said they still expected Senate approval of the complex and sweeping energy package, which would mark the first major overhaul of U.S. energy policy in more than a decade. <Frown A02>

Thus, in (1) the genitive NP under analysis is *the bill's supporters*, the possessor NP is *the bill*, the possessor NP's head noun is *bill*, and *bill* has a text frequency of 32 occurrences in Frown text A02 (which, like all texts under analysis, spans about 2000 words). Table 5 displays how thematicity of the possessor dovetails with genitive variation.

It is clear from Table 5 that Osselton's (1988) claim is correct: averaging over all corpora in our sample, the typical possessor head noun of an *s*-genitive has a text frequency of 5.94 occurrences while the typical possessor head noun of an *of*-genitive has a text frequency of only 3.88 occurrences, a difference which is highly significant at $p < .001$.

Two more specific findings strike us as remarkable: for one thing, *s*-genitives tend to be associated with significantly more frequent possessor head nouns in AmE than in BrE ($p < .001$). Secondly, while *s*-genitives come with significantly more frequent possessor head nouns in the 1990s than they do in the 1960s ($p < .001$), the reverse is true for *of*-genitive head nouns ($p < .001$). Frown exhibits these tendencies in an especially marked way.

5.1.3 *Information status*

As with many other alternation phenomena in the grammar of English, information status as a higher-level pragmatic factor has often been described as a significant determinant in genitive choice. Thus, according to the literature, if the possessor is given, the *s*-genitive is generally preferred because it places the given element first (see Biber et al. 1999b: 305–6; Quirk et al., 1985: 1282). To enable automatic coding of this factor, we chose to operationalize information status in a fairly straightforward way: for every possessor NP in our database, a Perl script established whether the head noun of the possessor NP occurred anywhere up to 50 words prior to the genitive slot

under analysis; if it did, the possessor phrase was classified as ‘given’. The example in (2) will illustrate:

- (2) The explosion sent the hood of the *car* flying over the roof of the house. The left front wheel landed 100 feet away. Police laboratory technicians said the explosive device, containing either TNT or nitroglycerine, was apparently placed under the left front wheel. It was first believed the bomb was rigged to *the car’s starter*. <BROWN A09>

In this passage, the genitive construction under analysis is *the car’s starter*; the possessor head noun is *car*, which had occurred 44 orthographic words prior to the genitive slot (. . . *the car flying over* . . .). Therefore, the possessor in this genitive token was classified as ‘given’. According to a univariate analysis of our database, information status indeed appears to be a determinant of genitive choice: while in our database as a whole, 26.9 percent of all *s*-genitive possessors are given, the corresponding figure for *of*-genitive possessor is only 17.6 percent, a difference which is highly significant ($p < .001$). Differences between sampling times or geographic differences are not statistically significant.

5.2 Phonological factors

5.2.1 Final sibilant in the possessor

Previous scholarship has shown that the presence of a final sibilant in the possessor, as in (3), may discourage the use of the *s*-genitive (cf. Altenberg 1982; Zwicky 1987):

- (3) But that is the sad and angry side of Bush. <Frown A11>

The phenomenon can be considered a phonological *horror aequi* effect (Rohdenburg 2000): because *Bush* ends in a sibilant (/S/), language users – according to the theory – avoid an immediately adjacent sibilant in the form of an *s*-genitive (i.e. *Bush’s sad and angry side*) and choose an *of*-genitive instead (which the writer did in (3)). We operationalized this phonological constraint, which also captures the effect that regular plural morphemes affixed onto the possessor NP have on genitive choice, by having a Perl script identify all possessors that end, orthographically, in <s> (as in *Congress*), <z> (as in *jazz*), <ce> (as in *resistance*), <sh> (as in *Bush*), or <tch> (as in *match*), and by coding all such possessors as ‘final sibilant present’.²² Table 6 gives the distributional results, by genitive type and corpus.

While the constraint is certainly not categorical, it does have the effect suggested in the literature: on the whole, only 12.0 percent of all *s*-genitives in our database have possessors that end in a final sibilant, while the corresponding percentage for *of*-genitives is 28.1 percent. This statistical skew is highly significant at $p < .001$. The constraint seems to be more powerful in our BrE data than in our AmE data: in Brown and Frown, 13.4 percent of all *s*-genitive possessors exhibit a final sibilant, but only 10.5 percent of all possessors in LOB and F-LOB do ($p < .05$). This skewing is mainly

²² Possessors ending in <dge> (as in *judge*) are so rare that they were excluded from analysis.

Table 6. *Presence of a final sibilant in the possessor by corpus and genitive type*

	<i>s</i> -genitive		<i>of</i> -genitive	
	%	<i>N</i>	%	<i>N</i>
Brown A/B	12.8%	102 (797)	29.4%	413 (1407)
LOB A/B	13.2%	100 (756)	27.2%	343 (1263)
Frown A/B	13.8%	156 (1134)	29.3%	292 (998)
F-LOB A/B	8.2%	73 (891)	26.6%	280 (1054)
Total	12.0%	431 (3578)	28.1%	1328 (4722)

due to F-LOB, which exhibits a particularly low percentage of *s*-genitives with final sibilant possessors.

5.3 Factors related to processing and parsing

This factor group subsumes all those constraints whose effects have been said to facilitate parsing (e.g. end-weight) or to avoid processing difficulties (e.g. nested genitives), or factors which can be (partly) explained by properties of the human speech production system (e.g. persistence).

5.3.1 End-weight

According to the time-honored principle of ‘end-weight’ (for instance, Behaghel 1909/10; Wasow 2002), language users tend to place ‘heavier’, more complex constituents after shorter ones, which yields a constituent ordering that might facilitate parsing (see, for example, Hawkins 1994). It has been claimed that the principle of end-weight impacts on the alternation between the *s*-genitive and the *of*-genitive as follows: 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, we expect a general preference for the *s*-genitive because it places the possessum last (Altenberg 1982: 76–9; Biber et al. 1999b: 304–5; Kreyer 2003: 200–4; Quirk et al. 1985: 1282; Rosenbach 2005; among many others). For the purposes of the present study, we sought to approximate the weight of genitive constituents by determining their length in graphemic words, utilizing Perl scripts for automatic coding. For illustration, consider (4):

- (4) Latter domain, under *the guidance of Chef Tom Yokel*, will specialize in steaks, chops, chicken and prime beef as well as Tom’s favorite dish, stuffed shrimp. <Brown A31>

In the *of*-genitive in (4) (*the guidance of Chef Tom Yokel*), the possessor phrase consists of three words (*Chef Tom Yokel*) and the possessum of two words (*the guidance*). Note, however, that if the writer had opted for an *s*-genitive instead, the possessum phrase could not have been determined by an article (**Chef Tom Yokel’s the guidance*). Therefore, definite or indefinite articles determining the possessum phrase of an *of*-genitive were not included in the tally in order not to skew results (see Altenberg 1982:

Table 7. *Mean possessor length by corpus and genitive type*

	<i>s</i> -genitive		<i>of</i> -genitive	
	mean	std. dev.	mean	std. dev.
Brown A/B	1.83	.75	2.83	1.56
LOB A/B	1.88	.74	2.55	1.29
Frown A/B	1.78	.83	2.60	1.38
F-LOB A/B	1.79	.79	2.83	1.46
Total	1.81	.78	2.71	1.44

79–84 for a similar coding procedure). Thus, net possessum length of the possessum phrase in (4) is exactly one word (*guidance*). In all, given the 1:3 ratio of possessum-to-possessor length, the use of the *of*-genitive in (4) is expected along the lines of the principle of end-weight.²³

Let us now turn to an overview of mean possessor lengths in our sample, as usual by corpus and genitive type (table 7). Mean length of *s*-genitive possessors is 1.81 words while mean length of *of*-genitive possessors is 2.71, a difference which is highly significant ($p < .001$). As expected, given the literature on end-weight, the *of*-genitive is thus preferred with longer possessors because it places the possessor phrase last. Relative to that, the *s*-genitive prefers shorter possessors, and has been doing so increasingly: mean *s*-genitive possessor length in the 1960s is 1.86, but 1.78 in the 1990s ($p < .01$). We also want to draw attention to the fact that the standard deviation associated with *s*-genitive possessor length is only roughly half of what it is for *of*-genitive, which is another way of saying that the *s*-genitive is more consistently restricted to short possessors than the *of*-genitive is to longer possessors.

According to the logic of the end-weight principle, the distribution of mean possessum lengths should be a mirror image of the distribution of mean possessor lengths. This is indeed the case, according to table 8: with a mean net length of 1.76, *s*-genitive possessums are on the whole longer than *of*-genitive possessums (1.51). While the difference in mean possessum net lengths by genitive type is not as marked as the difference in mean possessor length, it is still highly significant at $p < .001$.

At the same time, the data exhibit significant geographic and longitudinal differences: for one thing, at a mean length of 1.67, AmE on the whole prefers longer possessums than BrE, where mean length is 1.56 ($p < .001$). By contrast, while the *of*-genitive has

²³ These technical issues aside, we would like to stress that we do not want to claim for a minute that length of a phrase in words and heaviness of that phrase are exactly the same thing. Rather, we utilize length as a proxy for weight, a method that – besides having a tradition in the study of weight effects in genitive choice (e.g. Altenberg 1982; Kreyer 2003; Rosenbach 2005) – does strike us as rather unproblematic: Wasow (1997) concluded that ‘it is very hard to distinguish among various structural weight measures as predictors of weight effects. Counting words, nodes, or phrasal nodes all work well’ (1997: 102), and Szmrecsanyi (2004) demonstrated statistically that counting words is an excellent way of approximating syntactic node counts as a measure of syntactic complexity.

Table 8. *Mean possessum length (net) by corpus and genitive type*

	<i>s</i> -genitive		<i>of</i> -genitive	
	mean	std. dev.	mean	std. dev.
Brown A/B	1.81	1.13	1.57	.87
LOB A/B	1.63	.97	1.52	.87
Frown A/B	1.84	1.16	1.52	.84
F-LOB A/B	1.74	1.02	1.43	.71
Total	1.76	1.08	1.51	.83

come to prefer shorter possessors over time (mean *of*-genitive possessum net length was 1.55 in the 1960s, but is 1.47 in the 1990s; $p < .005$), the *s*-genitive is associated, in the 1990s, with longer possessors than it was in the 1960s (1.79 vs. 1.72; $p < .05$).

5.3.2 Persistence

We now move on to a further processing-related constraint on genitive choice, viz. precedence of an identical genitive construction in the preceding textual discourse. Psycholinguists, discourse analysts, and corpus linguists alike have amassed ample evidence that language users tend to reuse linguistic material that they have used or heard before; depending on the analytical perspective, this phenomenon has been called ‘persistence’ (e.g. Szmrecsanyi 2006), ‘priming’ (e.g. Bock 1986), ‘structural parallelism’ (e.g. Weiner & Labov 1983), or simply ‘repetition in discourse’ (e.g. Tannen 1989). It is known that persistence – the term we will adopt – significantly impacts on genitive choice in spoken English (Szmrecsanyi 2006: 87–101), and given that the phenomenon is known to be observable in both spoken and written language (cf. Gries 2005) we expect to see a persistence effect in our written data sample as well. The idea in a nutshell is that usage of, say, an *s*-genitive in a given genitive slot increases the odds that the writer will use an *s*-genitive again next time she has a choice (which is another way of saying that the share of *s*-genitives preceded by another *s*-genitive should be significantly higher than the share of *of*-genitives preceded by an *s*-genitive). In this spirit, we relied on Perl scripts to establish, for each genitive occurrence in our database, whether an *s*-genitive had been used last time there was a genitive choice. Example (5) illustrates a context where two subsequent interchangeable genitive contexts (*the continent’s river systems* and *the country’s Medical Association*) both exhibit *s*-genitives:

- (5) In both countries the cases appeared to indicate what is most feared: that *the continent’s river systems* are now infected, making the spread of the disease extremely difficult to control. In Ecuador, *the country’s Medical Association* said 100 people had died of a total of 5000 cases. . . <F-LOB A14>

The working hypothesis is borne out by our data: 50.4 percent of all *s*-genitives in our sample are preceded by another *s*-genitive, but only 37.5 percent of all *of*-genitives; this difference is statistically highly significant ($p < .001$). Compared

to the 1960s, significantly more genitives overall are preceded by an *s*-genitive in the 1990s ($p < .001$), an effect which is partly due to the fact that, as we have seen, the overall number of *s*-genitives in the data has increased significantly as well.

5.3.3 *Nested genitives*

Nested genitives are yet another phenomenon that broadly falls into the category of language users avoiding structures that are presumably difficult to parse or process. More specifically, our hypothesis is that (i) the *s*-genitive is preferred when either the possessor or possessum contains a nested *of*-genitive, and that (ii) the *of*-genitive is preferred when either the possessor or possessum contains a nested *s*-genitive. In other words, we posit a morphosyntactic *horror aequi* effect (cf. Rohdenburg 2000) such that language users avoid two identical genitive constructions in the same NP.

To corroborate the effect empirically, a Perl script identified all nested genitives in our database. We first turn to nested *s*-genitives, as in (6):

- (6) Because of [[*the recent death*]_{possessum} of [*the bride's father*]_{possessor}]. . . the marriage of Miss Terry Hamm to John Bruce Parichy will be a small one at noon tomorrow in St. Bernadine's church Forest Park. <Brown A16>

In (6), the possessor phrase contains a nested *s*-genitive (*the bride's father*), which might be one reason why the superordinate genitive construction is realized as an *of*-genitive and not as an *s*-genitive. Note here that *the bride's father's recent death*, besides not being very aesthetic, is probably also more difficult to parse because it is demonstrably more complex internally, the scope of each 's'-marker being slightly opaque. A univariate analysis of our database suggests that *s*-genitives with nested *s*-genitives are a rare phenomenon indeed: only 0.4 percent of the *s*-genitives in our database command another nested *s*-genitive, while 4.4 percent of all *of*-genitives in our database do (needless to say, this difference is highly significant at $p < .001$). On the whole, however, nested *s*-genitives are quite rare, though it seems worth pointing out that *of*-constructions with nested *s*-genitives, precisely as in (6), have become significantly more frequent over time ($p < .005$). This type of nesting is a particularly efficient kind of noun-phrase internal information packaging, and therefore anticipates the discussion in section 5.4 below on economy-related factors.

Example (7) provides an example of an *of*-construction – though not an interchangeable one (*the representatives' house* would not identify the institution) – that is nested into a superordinate *s*-genitive construction:

- (7) Also in [[*the House of Representatives*']_{possessor} [*bill*]_{possessum}] was more than \$65 million for refurbishing the Presidio over the next two years. <Frown A02>

Recall that our hypothesis was that a higher percentage of *s*-genitives than of *of*-genitives in our database would exhibit a nested *of*-genitive. This hypothesis cannot be corroborated by a univariate analysis of our database as a whole; as for individual corpora, only Frown displays the theoretically expected skewing, albeit not in a statistically significant fashion.

5.4 Economy-related factors

5.4.1 Type-token ratio

Szmrecsanyi (2006: 97) has shown that in spoken language, speakers prefer the *s*-genitive in contexts characterized by high type–token ratios, which he took to be indicative of increased lexical density. Along similar lines, we propose that the *s*-genitive is preferred by writers in contexts where lexical density is high, and thus where there is a need to economically code more information in a given textual passage. The rationale for this claim is that the *s*-genitive can be seen as the more compact and thus economic coding option vis-à-vis the *of*-genitive, which, according to Biber et al. ‘produces a less dense and more transparent means of expression’.²⁴ The *s*-genitive, by contrast, ‘represents a good way of compressing information’ (1999b: 302).²⁵

In order to test our claim that greater preference for the *s*-genitive correlates with lexical density, and to thus present a dynamic perspective on Biber’s claim that increasing density of information yields higher information load in noun phrases (2003), we utilized Perl scripts to establish the type–token ratios of the textual passages where the genitive occurrences in our database are embedded (i.e. 50 words before and 50 words after a given genitive construction). In a similar vein as Szmrecsanyi (2006), then, we consider type–token ratio a proxy variable for lexical density: the more different word types we find in a given passage, the higher the lexical density and the more pressing the need to code economically. Let us illustrate our coding procedure with a concrete example: (8) exhibits a genitive slot (*the miseries of families*) embedded in a passage with one of the highest type–token ratios in the entire Frown corpus: 89 different word types in a passage of about 100 words:

- (8) ...ridicule Dukakis.) So in 1992, by Quayle’s interesting subliminal design, Murphy carries at least some of Willie’s message: mindless liberalism allied with black anarchy (ruined families, unwed mothers, crime, drugs) leads quickly to social breakdown. If Quayle has no malign racial-political intent, he might point out, when discussing *the miseries of families*, that, for example, Eastern prep schools are filled with children packed off to get them away from divorce, incest, alcoholism, child abuse, wife battering and other horrors at home. The willingness to let the racist implication stand unchallenged, unexamined, loitering on the threshold, is the ugliest aspect of all this.
<Frown A12>

Given this high type–token ratio, the fact that the genitive construction under analysis is coded with an *of*-genitive, and not with an *s*-genitive, is unexpected. What about the database as a whole? Table 9 shows that in general, *s*-genitives are indeed associated with higher type–token ratios (the mean value is 74.9) than *of*-genitives, which yield a mean value of 73.0 (a difference which is statistically highly significant at $p < .001$). Notwithstanding this finding – which squares with our research hypothesis – there

²⁴ Likewise, Barber (1964: 132–3) and Potter (1969: 105) pointed out earlier that the *s*-genitive is the more concise, compact, and therefore economical choice of the two.

²⁵ As Raab-Fischer (1995: 124) notes, this may well be the reason why *s*-genitives are more frequent in newspaper language than in general usage, as has been noted (Dahl 1971; Jahr Sorheim 1980).

Table 9. *Mean type–token ratio of text passage where genitive slot is embedded by corpus and genitive type*

	<i>s</i> -genitive		<i>of</i> -genitive	
	mean	std. dev.	mean	std. dev.
Brown A/B	75.2	5.1	73.2	5.3
LOB A/B	70.7	5.0	69.0	4.7
Frown A/B	76.3	5.1	75.0	5.1
F-LOB A/B	76.4	4.5	75.4	4.7
Total	74.9	5.4	73.0	5.6

are differences between the two varieties and periods tested: AmE genitive contexts typically have a higher type–token ratio than BrE genitive contexts (74.8 vs. 72.7; $p < .001$), and genitive contexts in our 1990s subsample score considerably higher values than genitive contexts in our 1960s subsample (75.8 vs. 71.9; $p < .001$).

5.4.2 ‘Nouniness’

Another way of tapping into economy-related constraints, we suggest, is to assess a given passage’s ‘nouniness’, that is, the number of nouns it exhibits. The idea is that increased ‘nouniness’ – much in analogy to increased lexical density – is indicative of a local need to code, in a given textual passage, as much information as possible, whereas higher frequencies of verbs are found in more involved, narrative, and colloquial genres (Biber 2003: 179; Mair et al. 2002: 255–6). Under such circumstances, the *s*-genitive should be the preferred option due to its relative economy and its more ‘nouny’ structural design (cf. Biber et al. 1999b: 300–2). Thus, we had Perl scripts count the number of words preceded by a nominal POS-tag (<w N*>) in the textual passages (i.e. 50 words before and 50 words after a given genitive construction) where the genitive occurrences in our database are embedded. To illustrate, (9) gives one of the genitive passages in our database most packed with nouns (63 nouns in a textual snippet of about 100 words):

- (9) ...opening Saturday, June 3. Music for dancing will be furnished by Allen Uhles and his orchestra, who will play each Saturday during June. Members and guests will be in for an added surprise with the new wing containing 40 rooms and suites, each with its own private patio. Gene Marshall, *genial manager of the club*, has announced that the Garden of the Gods will open to members Thursday, June 1. Beginning July 4, there will be an orchestra playing nightly except Sunday and Monday for the summer season. Mrs. J. Edward Hackstaff and Mrs. Paul Luetete are planning a luncheon next week in honor... <Brown A17>

A univariate analysis of this factor suggests that in accordance with our working assumption, the average *s*-genitive passage, with a mean value of 28.4, is indeed ‘nounier’ than the typical *of*-genitive passage with a mean value of 27.6 ($p < .001$). Again, though, there is a significant difference between the two varieties: AmE genitive

passages, with a mean value of 28.8, are typically more ‘nouny’ than BrE genitive passages, which score a mean value of 27.0 ($p < .001$).

6 Multivariate analysis

We will now use *binary logistic regression* (see Pampel 2000 for an introduction) to quantify the combined contribution of all the conditioning factors discussed so far. As a multivariate procedure, logistic regression integrates probabilistic statements into the description of performance and is applicable ‘wherever a choice can be perceived as having been made in the course of linguistic performance’ (Sankoff & Labov 1979: 151). At base, logistic regression predicts a binary outcome (i.e. a linguistic choice) given several independent (or predictor) variables, thus having the following advantages over more simple, univariate analysis methods: (i) logistic regression estimates the effect size of each predictor; (ii) it specifies the direction of the effect of each predictor; (iii) it quantifies how much of the empirically observable variance is explained by the predictors considered; (iv) it states how well the model fares in predicting actual speakers’ choices; (v) it removes statistical artifacts and is invulnerable to epiphenomenal effects that may go undetected in univariate analysis. This last point is particularly important: in univariate analysis (for instance, when investigating animacy and end-weight separately), it is hard to determine whether two factors independently influence the outcome or whether one factor is epiphenomenal (because, for instance, animate possessors are typically shorter than inanimate possessors). Regression analysis, in point of fact, is the closest a corpus linguist can come to conducting a controlled experiment: the procedure systematically tests each factor while holding the other factors in the model constant.

In our analysis, we are going to report the following regression measures:

The magnitude and the direction of the influence of each predictor on the outcome.

This information is provided by *odds ratios*, indicating how the presence or absence of a feature (for categorical independents) or how a one-unit increase in a scalar independent influences the odds for an outcome. Odds ratios can take values between 0 and ∞ ; the more the figures exceed 1, the more highly the effect favors a certain outcome; the closer they are to zero, the more disfavoring the effect.

Model χ^2 . This measure tests the predictive ability of all the independents included in the model and indicates whether a model is statistically significant overall.

-2 log likelihood. This measure indicates how well the model fits the data. Smaller values are better than bigger values; a ‘perfect’ model has a -2 log likelihood value of zero.

Variance explained by (or explanatory power of) the model as a whole (R^2). The R^2 value can range between 0 and 1 and indicates the proportion of variance in the dependent variable (i.e. in the outcomes) accounted for by all the independent variables included in the model. Bigger R^2 values mean that more variance is accounted for by the model.

The specific R^2 measure which is going to be reported is the so-called *Nagelkerke R^2* , a pseudo R^2 statistic for logistic regression.

Predictive efficiency of the model as a whole. The percentage of correctly predicted cases vis-à-vis the baseline prediction (*% correct (baseline)*) indicates how accurate the model is in predicting actual outcomes. The higher this percentage, the better the model.

6.1 *The contribution of individual factors to genitive choice*

Given the conditioning factors described above, we estimated a logistic regression model on the combined samples of all four corpora subject to analysis (Brown, LOB, Frown, F-LOB). The model's dependent variable is the occurrence of an *s*-genitive instead of an *of*-genitive. Longitudinal, geographical, and genre differences were modeled by including three categorical variables as predictors and as moderators in interaction terms: TIME²⁶ (1990s vs. 1960s), VARIETY (AmE vs. BrE), and GENRE (B 'Editorials' vs. A 'Reportage'). Leaving interaction terms between internal constraints and external variables aside for a second (note that there were no significant interactions *between* internal constraints),²⁷ table 10 reports odds ratios associated with the individual predictors included in the model, as well as some model summary measures. With regard to the latter, note that the model accurately predicts 79.1 percent of all outcomes and accounts for roughly half of the variance in the dependent variable ($R^2 = .510$).²⁸ The other half of the variance may be due to semantic factors proper, to non-semantic factors not considered in this study, or to free variation.²⁹

As for effect sizes, almost all of the predictors included in the regression model are significant and have the theoretically expected effect, given the literature; the only exception is GIVENNESS OF THE POSSESSOR HEAD, which is not selected as significant (though note that if the end-weight-related predictors and thematicity of the possessor head³⁰ are removed from the model, GIVENNESS is selected as actually significant – which

²⁶ In what follows, predictors in logistic regression will appear in small capitals.

²⁷ Specifically, we tested for an interaction effect between possessor length and possessum length: a corresponding interaction term is not selected as significant in regression analysis, and does not add substantially to the model's overall explanatory power.

²⁸ A brief illustration of the difference between 'variance' (a rather abstract notion which measures the statistical dispersion of the dependent variable) and 'percentage of correctly predicted outcomes' might be helpful here. Consider a 'dumb' model which categorically predicts the *of*-genitive: this model would correctly predict 56.9 percent of the outcomes (because this is the overall percentage of *of*-genitives in our dataset, cf. table 2), but it would, sure enough, explain none of the statistical variance between the *of*-genitive and the *s*-genitive in our data.

²⁹ N is not 8,300, because a number of cases with missing values were excluded from regression analysis. This concerns the first genitive instance, which cannot have a genitive precedence, in each of the 284 corpus texts under analysis.

³⁰ Givenness and thematicity are somewhat correlated in that a highly thematic possessor NP is inherently given. It may also be worth pointing out that while the 50-word window (cf. section 5.1.3) that we used to establish whether a possessor NP is given is, admittedly, quite arbitrary, experimentation with other textual windows (including smaller ones) did not add to the significance level of GIVENNESS.

Table 10. *Logistic regression estimates: individual genitive predictors. Predicted odds are for the s-genitive*

	odds ratio
ANIMACY OF POSSESSOR	***
collective vs. inanimate	4.44 ***
animal vs. inanimate	7.09 *
human vs. inanimate	13.93 ***
LN OF TEXT FREQUENCY OF POSSESSOR HEAD	1.18 **
GIVENNESS OF POSSESSOR HEAD	1.09
FINAL SIBILANT IN POSSESSOR	.34 ***
LENGTH OF THE POSSESSOR PHRASE	.41 ***
LENGTH OF THE POSSESSUM PHRASE	.99
PERSISTENCE	1.15 *
PRESENCE OF NESTED S-GENITIVE	.33 ***
PRESENCE OF NESTED OF-GENITIVE	2.91 ***
TTR OF THE EMBEDDING PASSAGE (1 unit = 10 points)	1.82 ***
'NOUNINESS' OF THE EMBEDDING PASSAGE (1 unit = 10 points)	1.09 *
GENRE (B vs. A)	.68 *
VARIETY (AmE vs. BrE)	.69
TIME (1990s vs. 1960s)	.80
model intercept	.01***
<i>N</i>	8,015
<i>model</i> χ^2	3791.78 (<i>df</i> =31) ***
<i>-2 log likelihood</i>	7163.76
<i>Nagelkerke R²</i>	.506
<i>% correct (% baseline)</i>	79.1 (57.0)

*significant at $p < .05$, **significant at $p < .01$, ***significant at $p < .005$

strongly suggests that information status is indeed epiphenomenal to other factors such as weight along the lines of Hawkins 1994).

VARIETY and TIME as main effects are not selected as significant, hence geography and sampling time do not *per se* have an effect on genitive choice (though we will see later that VARIETY and TIME interact significantly with a number of internal variables). GENRE, by contrast, does have a *per se* effect: finding a genitive slot in the B section (in editorials, that is, instead of in the reportage section) significantly decreases the odds for an *s*-genitive by 33 percent.³¹

Figure 4 provides a quantitative comparison of the dichotomous internal factors in our variable portfolio. The first three factors in this diagram are ANIMACY predictors. It is striking here how felicitously Rosenbach's (2006: 105) animacy hierarchy (human >

³¹ This finding squares with accounts (e.g. Mair et al. 2002) that reportage is leading other genres in the colloquialization of the written norm if, like Leech & Smith (2006), we tentatively regard the increase of the *s*-genitive as a colloquialization process; see however our discussion of this claim in the conclusion.

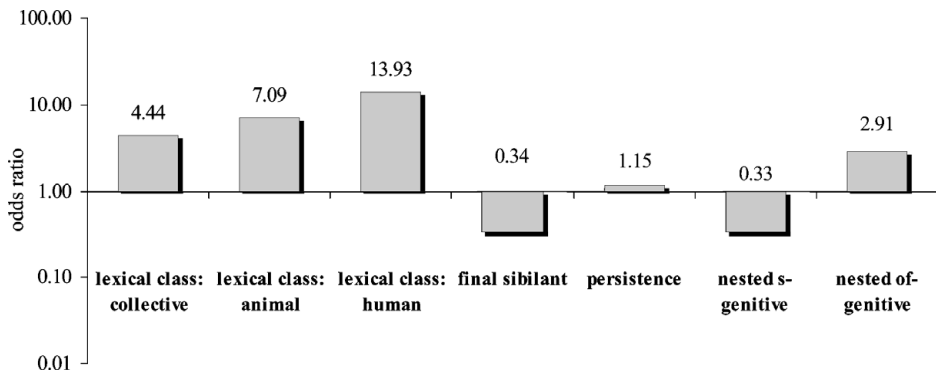


Figure 4. Odds ratios associated with categorical, internal factors in logistic regression (predicted odds are for the *s*-genitive)

animal > collective > inanimate) is rendered by our data set: if the possessor head noun is a collective noun instead of an inanimate noun (this being the baseline condition), the odds for the *s*-genitive increase by a factor of 4.51; if we are dealing with an animal possessor, the odds for an *s*-genitive increase about eight-fold; and if the possessor head noun is a human noun, the odds for an *s*-genitive increase almost fourteen-fold. A human possessor is thus the single most powerful categorical predictor in our variable portfolio. A FINAL SIBILANT in the possessor decreases the odds for an *s*-genitive by 66 percent; precedence (PERSISTENCE) of an *s*-genitive in the last slot increases the odds for an *s*-genitive in the slot under analysis by 15 percent; and a nested *s*-genitive decreases the odds for an *s*-genitive in the superordinate syntagm by 61 percent, while, conversely, a nested *of*-genitive increases the odds for an *s*-genitive by a factor of 3.64. The direction of all of these effects is as expected.

As for the scalar (i.e. nondichotomous) predictors in our analysis, note that it would not be meaningful to arrange them in a comparative diagram akin to figure 4 since scalar predictors are not necessarily on the same scale (a one-word increase in, for example, POSSESSOR LENGTH is not really equivalent to a one-unit increase in, say, TYPE-TOKEN RATIO). Still, the regression estimates in table 10 are instructive for gauging the relative importance of predictors. To start with, we modeled TEXT FREQUENCY OF THE POSSESSOR HEAD (i.e. ‘thematicity’ of the possessor along the lines of Osselton 1988) logarithmically to alleviate the effect of frequency outliers; thus, a text frequency of, for instance, 6 occurrences was rendered as $\ln(6) = 1.79$ in logistic regression. Given that, it turns out that for every one-unit increase in this measure (this corresponds to a frequency differential of, very roughly, 3 occurrences instead of 1 occurrence), the odds for an *s*-genitive increase by 17 percent; the direction of this effect is hence as hypothesized. Second, for every additional word in the POSSESSOR PHRASE, the odds for the *s*-genitive decrease by 74 percent, while every additional word in the POSSESSUM PHRASE increases the odds for an *s*-genitive about two-fold. Again,

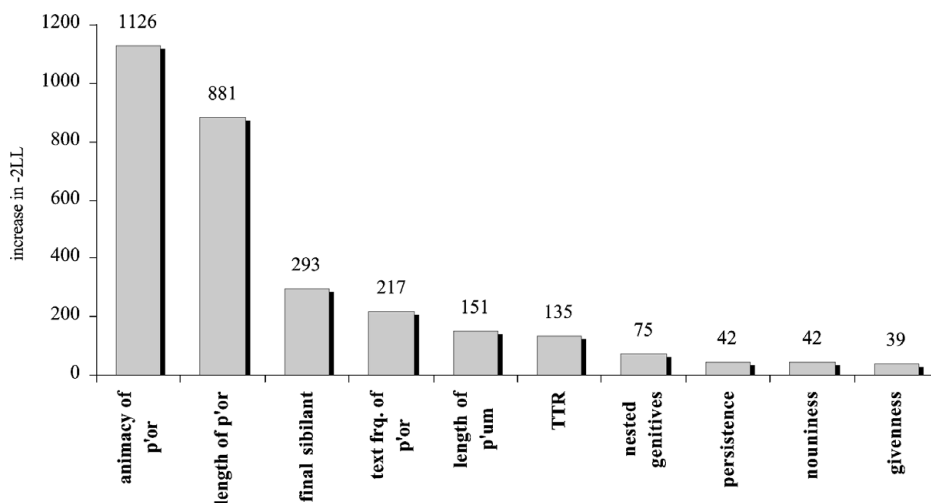


Figure 5. Increase in $-2 \log$ likelihood (decrease in model goodness-of-fit) if factor(s) removed

these effect directions are fully consonant with our research hypothesis. Going on to our economy-related predictors (TTR and ‘NOUNINESS’ OF THE EMBEDDING PASSAGE), we observe that every 10-point increase in the type-token ratio of a given genitive passage (that is, if we find, say, 60 different word types in a 100-word passage instead of just 50 different types) increases the odds for an *s*-genitive in that passage by 82 percent; for every 10 additional nouns in that passage, the odds for an *s*-genitive increase by 9 percent. While both of these measures thus have the theoretically expected effect direction – writers indeed prefer the *s*-genitive in lexically dense and ‘nouny’ contexts – lexical density turns out to be a substantially more important constraint than ‘nouniness.’

So far, we have been concerned with *effect sizes* – thus, we discussed how the outcome is affected, for instance, when a nested genitive is present. Crucially, *effect sizes* have to be distinguished from notions such as *explanatory power* of individual variables, and *goodness of fit* of the model as a whole. Thus, nested genitives may have a considerable effect size, as we have seen, when they are present, but it just so happens that nested genitives are very rare. Presumably, a model that simply ignores nested genitives would not lose too much of either its goodness of fit or its explanatory power.

With this in mind, we will now test the internal conditioning factors in our model to establish how crucial they are, from a bird’s eye perspective, for predicting genitive choice. Figure 5 ranks individual conditioning factors in our model in terms of how much they contribute to accounting for genitive choice in logistic regression. More specifically, figure 5 displays the increase in $-2 \log$ likelihood, a goodness-of-fit measure in logistic regression, when (groups of) factors – say, nested genitives – and interactions

with these factors are removed from the model.³² In a sense, figure 5 illustrates how much our model of genitive choice suffers when individual factors are omitted. Given these statistical technicalities, which factors are the really important ones in genitive choice? The single most crucial factor is ANIMACY OF THE POSSESSOR, followed – at some distance – by LENGTH OF THE POSSESSOR PHRASE. Observe, however, that the combined contribution of LENGTH OF THE POSSESSOR PHRASE and LENGTH OF THE POSSESSUM PHRASE is virtually identical to the contribution of ANIMACY. Thus, animacy and end-weight are the empirical pillars of the model, which is why our evidence lends strong empirical support to Rosenbach's claim (contra, e.g., Hawkins 1994) that the animacy effect 'cannot be reduced to an effect of weight (and vice versa)' (2005: 638). Indeed, we have seen that animacy has an independent effect on genitive choice, over and beyond the effect of the fact that animate possessors tend to be shorter than inanimate possessors. On the other hand, it is noteworthy that GIVENNESS OF THE POSSESSOR HEAD ranks last in figure 5 (and is, as we have seen, insignificant in logistic regression anyway). This squares with Gries' corpus-based observation of 'the complete overall irrelevance of . . . givenness' (2002: 26).

On the whole, consideration of the combined contribution of the four major factor groups considered in this study (semantic and pragmatic factors, phonology, factors related to processing and parsing, and economy-related factors) yields the following hierarchy of relevance:

(10) semantics/pragmatics ~ processing/parsing > phonology > economy

Still, it is important to point out that however secondary phonology and economy may be to animacy and end-weight, the former are still powerful enough to tip the balance in favor of either genitive type when end-weight and animacy are working against each other – for instance, in the case of short inanimate or long animate possessors.

6.2 Interaction effects

In regression analysis, interaction terms (cf. Jaccard 2001) are used to determine how strongly the influence of a particular independent variable (the 'focal' independent) depends on the value of a second independent variable (the 'moderator' independent). The odds ratio associated with the interaction term is the multiplicative factor by which the main effect of the focal changes for a one unit increase (for scalar independents) or for a categorical coding (for dichotomous independents) of the moderator.

We will begin by discussing a genre effect in our data (table 11a). The odds ratio of 1.26 associated with the interaction term LENGTH OF THE POSSESSOR PHRASE * GENRE (B) indicates that for every one-word increase in the possessor phrase, the odds ratio

³² It would also have been possible to display the decreases in *Nagelkerke R²* instead. This would have yielded exactly the same ranking of factors. Note, along these lines, that the *-2 log likelihood* figures do not have an interpretation in absolute terms: what is important is the ranking they yield.

Table 11. *Logistic regression estimates: selected interaction terms (only significant interaction terms are displayed). Predicted odds are for the s-genitive*

	odds ratio
<i>a. interactions involving</i> GENRE	
LENGTH OF THE POSSESSOR PHRASE * GENRE (B)	1.26***
<i>b. interactions involving</i> VARIETY	
ANIMACY OF POSSESSOR (inanimate) * VARIETY (AmE)	***
collective * VARIETY (AmE)	.75*
human * VARIETY (AmE)	.54***
LN OF TEXT FREQUENCY OF POSSESSOR HEAD * VARIETY (AmE)	1.20**
LENGTH OF THE POSSESSUM PHRASE * VARIETY (AmE)	1.38***
<i>c. interactions involving</i> TIME	
LN OF TEXT FREQUENCY OF POSSESSOR HEAD * TIME (1990s)	1.40***
FINAL SIBILANT IN POSSESSOR * TIME (1990s)	.72*
LENGTH OF THE POSSESSOR PHRASE * TIME (1990s)	.65***
<i>d. interactions involving</i> TIME * VARIETY	
LENGTH OF THE POSSESSOR PHRASE * TIME (1990s) * VARIETY (AmE)	1.40***

*significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$

comparing the predicted odds for an *s*-genitive in B texts with the predicted odds for an *s*-genitive in A texts changes by a multiplicative factor of 1.26. Because the main effect of LENGTH OF THE POSSESSOR PHRASE is 0.41 (cf. table 10), the actual effect of the predictor in B texts is $0.41 \times 1.26 = 0.52$. This is another way of saying that length of the possessor phrase, and hence end-weight, is a less important factor in B texts (Press: Editorials) than it is in A texts (Press: Reportage). As a tentative explanation, we suggest that parsing efficiency may be a more pressing concern in reportage texts than in editorials, where other factors, such as stylistic constraints, may play a bigger role. Also, as noted above, the ‘Reportage’ genre has been shown to be particularly susceptible to colloquialization as an instantiation of popularization, and surely parsing efficiency can be considered a phenomenon indicative of more colloquial genres.

Next, how do differences between AmE (Brown, Frown) and BrE (LOB, F-LOB) play out in logistic regression? Consider table 11b: we obtain, for one thing, interaction terms between ANIMACY categories and VARIETY. Thus, while the main effect of collective nouns and human nouns is 4.44 and 13.93, respectively (cf. table 10), in AmE these categories yield values of 4.44×0.75 (3.33) and 13.93×0.54 (7.52), respectively. In short, this means that in AmE the effect of more animate possessors on the odds that an *s*-genitive will be chosen is significantly more moderate than in BrE. By the same token, less animate possessors discourage the *s*-genitive less forcefully in AmE than in BrE.

Second, the odds ratio of 1.20 associated with the interaction term LN OF TEXT FREQUENCY OF POSSESSOR HEAD * VARIETY shows that while the main effect of ‘thematic genitives’ favors the *s*-genitive with a factor of 1.18, the effect in our AmE dataset

specifically is 1.18×1.20 (1.42). Therefore, in AmE, for every one-point increase in this predictor, the odds for *s*-genitive increase by 42 percent, instead of just 18 percent, which means that thematicity is a substantially more crucial factor in our AmE data than in our BrE data.

Third, we observe a significant interaction between LENGTH OF THE POSSESSUM PHRASE and VARIETY; earlier, we detailed (cf. table 10) that LENGTH OF THE POSSESSUM PHRASE is not selected as a significant main effect in regression. What we observe now is that, unlike in our database as a whole, the predictor LENGTH OF THE POSSESSUM PHRASE is actually significant and has the theoretically expected effect in the AmE data. Hence, in Brown and Frown for every one-word increase in the possessum phrase, the odds for the *s*-genitive increase by a factor of 0.99×1.38 (1.37), i.e. by 37 percent. By contrast, length of the possessum phrase is irrelevant to genitive choice in the BrE data.

We now turn to a discussion of the diachronic trends in our data: the differences between data sampled in the 1960s (Brown, LOB), on the one hand, and data sampled in the 1990s (Frown, F-LOB), on the other hand (Table 11c).

First of all, observe that the factor ANIMACY does not interact significantly with sampling time when other factors, such as end-weight, are controlled for. Hence, whatever the longitudinal spread of the *s*-genitive in our data is due to, it does not seem to involve shifts in writer's preferences concerning animacy of the possessor. With regard to actually significant interactions, we saw above that in AmE, 'thematic genitives' (Osselson 1988) are more important than in BrE. Observe, now, that according to the interaction term LN OF TEXT FREQUENCY OF POSSESSOR HEAD * TIME in table 11c, this predictor also exhibits a longitudinal difference: in our 1990s corpora, every one-unit increase in the measure increases the odds for an *s*-genitive by a considerable 65 percent, instead of just 18 percent in our database as a whole. The emerging pattern, an increase in the importance of 'thematicity' as a factor that favors the inflected genitive, is thus clearly a case of Americanization, as the drift is led by AmE, with BrE trailing a little behind.

Second, the odds ratio of 0.65 associated with the term FINAL SIBILANT IN POSSESSOR * TIME indicates that the effect of the presence of a final sibilant had a substantially bigger magnitude in the 1990s than in the 1960s: whereas in our total database, the presence of this phonological condition decreases the odds for an *s*-genitive by 66 percent (cf. table 10), the corresponding figure for our 1990s subsample is 78 percent. It is somewhat paradoxical that a phonological constraint should become more powerful, over time, in written newspaper language – this can only be interpreted, we believe, as a type of 'colloquialization of the written norm' (cf. Hundt & Mair 1999, explained in section 1.2.2 above).

Third, longer possessor phrases disfavored the *s*-genitive more markedly in the 1990s than in the 1960s (LENGTH OF THE POSSESSOR PHRASE * TIME): the main effect of possessor length is 0.41 (cf. table 10), but in our 1990s subsample the constraint is associated with an odds ratio of 0.41×0.65 , hence 0.27 (which means that in Frown and F-LOB every additional word in the possessor phrase decreases the odds for an *s*-genitive by

73 percent). According to table 11d and the three-way interaction to be found there, though, what we just saw is truer for F-LOB than in Frown, where the effect of long possessors is, indeed, not out of the ordinary. In a word: in our F-LOB data, long possessors disfavor the *s*-genitive in a somewhat extreme fashion.

7 Summary and conclusion

In this section, we seek to assess our previous findings in terms of why the *s*-genitive has been spreading over time, and why this tendency has been more marked in AmE press material than in BrE press material.

7.1 *Why has the s-genitive been spreading in press texts?*

Recall that usage of the *s*-genitive has increased by 10 percentage points in the BrE data and by 16 percentage points in AmE data (see figure 3). Which factor(s) are responsible for this increase? While our univariate analysis suggested that the *s*-genitive has come to be associated with more inanimate possessors over time, an animacy effect could not be substantiated in multivariate analysis (that is, when other factors such as possessor length were controlled for). Somewhat surprisingly, then, our analysis suggests that the spread of the *s*-genitive especially in BrE is unlikely to be due to changes in the effect that possessor animacy has on genitive choice.

Also, the context we analyzed does not warrant the claim that NPs in general have become more animate over time (thus increasing the frequency of *s*-genitives, even with all other things being equal): a random sample of 2000 NPs in the four corpora did not yield any remotely significant differences in mean NP animacy between our 1960s and 1990s data. We offer, instead, as one of the reasons why the *s*-genitive might be on the increase, that ‘thematic’ possessor NPs (cf. Osselton 1988) – that is, NPs that have a high text frequency in a given corpus text – favor the *s*-genitive substantially more strongly in our 1990s data than in our 1960s data (see table 10). Along these lines, we should also add here that while we had at the outset classified thematic genitives as a pragmatic phenomenon, the factor can also be seen as an economy-related constraint: when writing about a noun or NP repeatedly, why not just as well use the economical *s*-genitive with that noun or NP?

Turning to yet another economy-related factor, recall that for every 10-point increase in a given genitive passage’s type–token ratio, the odds for the *s*-genitive increase by about 80 percent (cf. table 10), we argued that writers prefer the more compact coding option in lexically more dense environments. While regression analysis indicated that the effect size of this factor has stayed fairly constant over time, corpus texts in general seem to have become lexically more dense: supplemental analyses suggest that there has been a highly significant tendency over time, in both varieties, towards increased lexical density (mean number of different types per corpus text in the 1960s: 821.08; 1990s: 848.94; $p < .001$), a development which inherently favors the *s*-genitive. This is a circumstance that is *per se* unrelated to the system of genitive choice, but which indirectly favors the *s*-genitive as the more compact coding option.

In all, we find that the overall spread of the *s*-genitive in press language is not due to changes in the way animacy or end-weight constrain genitive choice, but may well reduce, at least partly, to an increasingly powerful tendency to code thematic NPs with the *s*-genitive, as well as to an epiphenomenon effect of an increasing overall lexical density of journalistic prose – a factor which would always have favored the *s*-genitive.

7.2 *Why has the s-genitive become so much more frequent in AmE press material than in BrE press material?*

In Frown, interchangeable *s*-genitives are 7 percentage points more frequent than in F-LOB. This differential is remarkable since Brown and LOB exhibit virtually the same share of interchangeable genitives. How can our analysis account for this divergence? First, animacy is an overall weaker factor in the AmE data than in the BrE data: our univariate analysis (see table 4) has shown that *s*-genitive possessors in Frown are significantly less animate than *s*-genitive possessors in F-LOB. In a similar vein, our regression estimates (see table 11b) indicated that inanimate possessors discourage the *s*-genitive less forcefully in AmE than in BrE. In other words, it is particularly in AmE that the *s*-genitive has spread with inanimate possessors, much more so than in BrE.

Secondly, logistic regression has shown that ‘thematic NPs’ (cf. Osselton, 1988) favor the *s*-genitive significantly more strongly in AmE than in BrE – thus, when choosing a genitive construction for a frequent, and thus more thematic, possessor, American journalists are significantly more likely to opt for an *s*-genitive than are their British counterparts, a preference which skews distributions in our AmE data in favor of the *s*-genitive.

Third, while length of the possessum phrase is not a significant factor in genitive choice for British journalists, we saw (Table 11b) that the factor is in fact significant for genitive choice in our AmE material: every additional word in the possessum phrase increases the odds for the *s*-genitive by 37 percent in Brown and Frown. Irrelevant as it is in the BrE data, this is another factor that systematically favors the *s*-genitive in the AmE data. In this context, recall also that we have seen (table 11d) that specifically in F-LOB, longer possessor phrases disfavor the *s*-genitive in an extreme fashion – a constraint that skews proportions in F-LOB in favor of the *of*-genitive. Two further characteristics of our AmE material, albeit unrelated *per se* to genitive choice, are nonetheless likely to also be responsible for the high frequency of the *s*-genitive especially in Frown. For one thing, additional analyses indicate that lexical density in general is higher in AmE texts (mean value: 845.84 different types per text) than in BrE texts (mean value: 822.65 types per text; $p < .001$); crucially, high type–token ratios favor, as we have seen, the *s*-genitive. On the other hand, we detailed earlier that frequent, and thus ‘thematic’, NPs are especially likely to be coded with the *s*-genitive in AmE press material. According to a random sample of 20,000 nouns taken from the four corpora, this effect is additionally amplified by the fact that the typical noun to be found in an AmE text has a significantly ($p < .005$) higher text frequency (mean value: 3.62 occurrences per text) than a noun occurring in a BrE text (3.37 occurrences).

In short, because AmE texts are more thematic at the outset, they exhibit more *s*-genitives.

To summarize, we have suggested that the *s*-genitive is more frequent in AmE press texts because (i) the *s*-genitive is less constrained by the factor animacy in AmE press texts, (ii) AmE writers are more likely to code frequent and thus thematic NPs with the *s*-genitive, (iii) AmE journalists, unlike their BrE counterparts, seem to consistently take into account the length of the possessum phrase (longer possessums favor *s*-genitive) while writers in F-LOB specifically abhor long *s*-genitive possessors (and thus opt more frequently for the *of*-genitive instead), and (iv) there are some textual characteristics of our AmE material – comparatively thematic (that is, textually frequent) nouns and high type–token ratios – that inherently favor the *s*-genitive.

7.2 Conclusion

Our multivariate analysis has shown that in the synchronic picture, genitive choice is dependent upon a complex mechanics of interlocking factors, no single one of which can be held solely responsible for the observable variation.

In the diachronic view we initially posed the question as to whether the continuing shift from *'s* to *of* can be described as an instance of colloquialization. Our results suggest that – given the unclear division of stylistic functions between *'s* and *of* – rather than a pure case of colloquialization, the case at hand is best explained as ‘economization’, i.e. as a response to the growing demands of economy, which, according to Biber (2003), are an ever-increasing force, particularly in newspaper language. Two central aspects of our findings support this assessment:

- (i) While the economy-related factor of textual density (see section 5.4.1) has not gained in explanatory power, the textual density of newspaper texts itself has increased significantly, following a typical Americanization pattern. This is, of course, a reflection of the ‘informational explosion’ (Biber 2003) that modern newspapers are faced with. Since the factor of textual density favors the *s*-genitive, it thus makes an important contribution to the diachronic shift in genitive variation.
- (ii) The factor that multivariate analysis has shown to have gained most dramatically in relative weight from the 1960s to the 1990s is ‘thematicity’ of the possessor head noun (see section 5.1.2). While we had good reason to treat it as part of our ‘semantic and pragmatic’ set of factors, it is obvious that this factor also relates to economy. If we understand ‘thematicity’ of a noun as a licensing factor for journalists to resort to the more compact *s*-genitive, it becomes plausible that in times of growing informational and textual density, writers should invoke this license more regularly. We showed that this increase of factorial weight for ‘thematicity’, like the increase in textual density, follows a pattern of Americanization.

In addition, we would like to point out that none of the factors that one might associate with colloquialization – e.g. those related to phonology or the semantics of animacy – could be shown to make a direct contribution to the diachronic shift in genitive variation;

recall also that a supplementary analysis failed to reveal a shift toward a more personal (i.e. less abstract) writing style.

On the methodological plane, we wish to emphasize, first, that the Brown series of corpora is ideally suited for large-scale – in terms of the number of cases studied – and yet sufficiently fine-grained quantitative research into frequent morphosyntactic phenomena such as genitive constructions. This is primarily due to the high overall quality of the POS-tagging in the dataset, which makes (semi-)automatic retrieval of the linguistic variable along with many of the relevant contextual parameters feasible. Second, this study has, we believe, demonstrated that the portfolio of factors conditioning (genitive) variation in time and space is best investigated by multivariate analysis methods. Conditioning factors partake, as we have seen, in a rather complex interplay with one another, and hence we could not agree more wholeheartedly with Anna Wierzbicka's observation that 'the overall picture produced by an analysis that pays attention to all the relevant factors is, admittedly, complex and intricate', yet it is 'the only kind of analysis that can achieve descriptive adequacy and explanatory power' (1998: 151).

What, then, is wrong with more traditional, univariate approaches to (genitive) variation – approaches, that is, which do not investigate factors simultaneously but one-by-one, usually relying on a series of crosstabulations? Crucially, univariate analysis methods may be unable to uncover significant effects, and are prone to fail 'in adequately describing, comprehensively explaining and successfully predicting' (Gries 2003: 185) linguistic variation. In particular, this means that whenever two or more factors in the variationist envelope are somewhat interrelated (as were, in our study, weight, animacy, and givenness of the possessor), univariate analysis techniques are, as a matter of fact, inappropriately reductionist and simplistic. This is why the present study might be viewed as an extended programmatic argument that whenever the set of independent variables exceeds a couple of (possibly not entirely independent) factors, corpus-based research into variation in time and space should adopt multivariate methodologies, which have long been state-of-the-art in variationist sociolinguistics and in the social sciences in general.

Authors' addresses:

Department of Linguistics

Stanford University

Margaret Jacks Hall

Stanford, CA 94304-2150

USA.

lhinr@stanford.edu

English Department

University of Freiburg

Rempartstr. 15

79098 Freiburg

Germany

benedikt.szmrecsanyi@anglistik.uni-freiburg.de

Appendix

Table A. *Text categories in the Brown family of matching 1-million-word corpora of written StE.*

Genre group	Category	Content of category	No. of texts
Press (88)	A	Reportage	44
	B	Editorial	27
	C	Review	17
General Prose (206)	D	Religion	17
	E	Skills, trades and hobbies	36
	F	Popular lore	48
	G	Belles lettres, biographies, essays	75
	H	Miscellaneous	30
Learned (80)	J	Science	80
Fiction (126)	K	General fiction	29
	L	Mystery and detective Fiction	24
	M	Science fiction	6
	N	Adventure and Western	29
	P	Romance and love story	29
	R	Humor	9
TOTAL			500

References

- Allen, C. L. 2003. Deflexion and the development of the genitive in English. *English Language and Linguistics* 7: 1–28.
- Altenberg, B. 1982. *The genitive v. the of-construction: A study of syntactic variation in 17th century English*. Malmö: CWK Gleerup.
- Barber, C. 1964. *Linguistic change in present-day English*. London and Edinburgh: Oliver and Boyd.
- Behaghel, O. 1909/10. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25.
- Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. 2003. Compressed noun-phrase structure in newspaper discourse: the competing demands of popularization vs. economy. In J. Aitchison & D. M. Lewis (eds.), *New media language*, 169–81. London and New York: Longman.
- Biber, D. & E. Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language* 65: 487–517.
- Biber, D. & E. Finegan. 2001. Diachronic relations among speech-based and written registers in English. In S. Conrad & D. Biber (eds.), *Variation in English: Multidimensional Studies*, 66–83. London: Longman.
- Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999a. *Longman grammar of spoken and written English*. Harlow: Longman.
- Biber, D., G. Leech & S. Johansson. 1999b. *Longman grammar of spoken and written English*. London and New York: Longman.
- Bock, K. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18: 355–87.

- Bresnan, J., A. Cueni, T. Nikitina & H. Baayen. Forthcoming. Predicting the dative alternation. In G. Boume, I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*. Amsterdam: Royal Netherlands Academy of Science.
- Burchfield, R. W. 1996. *Fowler's modern English usage*, 3rd edition. Oxford: Clarendon.
- Dahl, L. 1971. The *s*-genitive with non-personal nouns in modern English journalistic style. *Neuphilologische Mitteilungen* 72: 140–72.
- Denison, D. 1998. Syntax. In S. Romaine (ed.), *The Cambridge history of the English language*, vol. IV: 1776–1997, 92–329. Cambridge: Cambridge University Press.
- Dixon, R. M. W. 2005. *A semantic approach to English grammar*. Oxford: Oxford University Press.
- Fairclough, N. 1992. *Discourse and social change*. Cambridge: Polity.
- Fischer, O. 1992. Syntax. In N. Blake (ed.), *The Cambridge history of the English language*, vol. II: 1066–1476, 207–408. Cambridge: Cambridge University Press.
- Fischer, O. & W. Van Der Wurff. 2006. Syntax. In R. M. Hogg & D. Denison (eds.) *A history of the English language*, 109–98. Cambridge: Cambridge University Press.
- Fowler, H. W. 1926. *A dictionary of English usage*. Oxford: Oxford University Press.
- Francis, N. & H. Kučera 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
- Gries, S. T. 2002. Evidence in linguistics: three approaches to genitives in English. In R. M. Brend, W. J. Sullivan & A. R. Lommel (eds.), *LACUS Forum XXVIII: What constitutes evidence in linguistics*, 17–31. Fullerton, CA: LACUS.
- Gries, S. T. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York and London: Continuum.
- Gries, S. T. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34: 365–99.
- Hawkins, J. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
- Huddleston, R. & G. K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Hundt, M. & C. Mair. 1999. 'Agile' and 'uptight' genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221–42.
- Hundt, M., A. Sand & R. Siemund. 1998. Manual of information to accompany the Freiburg-LOB corpus of British English ('F-LOB'). Bergen: ICAME – International Computer Archive of Modern and Medieval English.
- Hundt, M., A. Sand & P. Skandera. 1999. Manual of information to accompany the Freiburg-Brown corpus of American English ('Frown'). Bergen: ICAME – International Computer Archive of Modern and Medieval English.
- Jaccard, J. 2001. *Interaction effects in logistic regression*. Thousand Oaks: Sage Publications.
- Jahr Sorheim, M.-C. 1980. *The s-genitive in present-day English*. Oslo: Department of English, University of Oslo.
- Jespersen, O. 1909–49. *A modern English grammar on historical principles*. Copenhagen: Munksgaard.
- Johansson, S., E. Atwell, R. Garside & G. Leech. 1986. *The tagged LOB corpus user's manual*. Bergen: Norwegian Computing Centre for the Humanities.
- Johansson, S. & K. Hofland 1989. *Frequency analysis of English vocabulary and grammar: based on the LOB corpus*. Oxford: Clarendon.
- Jucker, A. 1993. The genitive versus the *of*-construction in newspaper language. In A. Jucker (ed.), *The noun phrase in English: Its structure and variability*. Heidelberg: Carl Winter. 121–36.
- Kaye, A. S. 2004. On the bare genitive. *English Today* 20: 57–8.

- Kreyer, R. 2003. Genitive and *of*-construction in modern written English: Processability and human involvement. *International Journal of Corpus Linguistics* 8: 169–207.
- Krug, M. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: Mouton de Gruyter.
- Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715–62.
- Leech, G., B. Cruickshank & R. Ivanič. 2001. *An A-Z of English grammar & usage*, 2nd edition. Harlow: Pearson Education.
- Leech, G., B. Francis & X. Xu 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories In C. Fuchs & B. Victorri (eds.), *Continuity in linguistic semantics*, 57–76. Amsterdam and Philadelphia: John Benjamins.
- Leech, G. & N. Smith. 2006. Recent grammatical change in written English 1961–1992: Some preliminary findings of a comparison of American with British English. In A. Renouf & A. Kehoe (eds.), *The changing face of corpus linguistics*. 185–204, Amsterdam and New York: Rodopi.
- Mair, C. 2006a. Inflected genitives are spreading in present-day English, but not necessarily to inanimate nouns. In C. Mair (ed.), *Corpora and the history of English: Festschrift für Manfred Markus*. Heidelberg: Winter.
- Mair, C. 2006b. *Twentieth-century English: History, variation, and standardization*. Cambridge: Cambridge University Press.
- Mair, C., M. Hundt, G. Leech & N. Smith. 2002. Short-term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7: 245–64.
- Manning, C. D. 2003. Probabilistic syntax. In R. Bod, J. Hay & S. Jannedy (eds.), *Probabilistic linguistics*, 289–341, Cambridge, MA: MIT Press.
- Mustanoja, T. F. 1960. *A Middle English syntax*, part I. Helsinki: Société Néophilologique.
- Orwin, R. 1994. Evaluating coding decisions. In H. Cooper & L. Hedges (eds.), *The handbook of research synthesis*, 139–62. New York: Russell Sage Foundation.
- Osselton, N. 1988. Thematic genitives. In G. Nixon & J. Honey (eds.), *An historic tongue: Studies in English linguistics in memory of Barbara Strang*. London: Routledge.
- Pampel, F. 2000. *Logistic regression: A primer*. Thousand Oaks: Sage Publications.
- Peters, P. 2004. *The Cambridge guide to English usage*. Cambridge: Cambridge University Press.
- Potter, S. 1969. *Changing English*. London: Deutsch.
- Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language*. London and New York: Longman.
- Raab-Fischer, R. 1995. Löst der Genitiv die *of*-Phrase ab? Eine korpusgestützte Studie zum Sprachwandel im heutigen Englisch. *Zeitschrift für Anglistik und Amerikanistik* 43: 123–32.
- Rohdenburg, G. 2000. Implications of a horror aequi principle in Early and Late Modern English. *Proceedings of the Eleventh International Conference on English Historical Linguistics (11 ICEHL)*. Santiago de Compostela.
- Rosenbach, A. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin: Mouton de Gruyter.
- Rosenbach, A. 2003. Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In G. Rohdenburg & B. Mondorf (eds.), *Determinants of grammatical variation in English*, 379–412. Berlin and New York: Mouton de Gruyter.
- Rosenbach, A. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81: 613–44.
- Rosenbach, A. 2006. Descriptive genitives in English: A case study on constructional gradience. *English Language and Linguistics* 10: 77–118.

- Sand, A. & R. Siemund. 1992. LOB – 30 years on. . . *ICAME Journal* **16**: 119–22.
- Sankoff, D. & W. Labov. 1979. On the use of variable rules. *Language in Society* **8**: 189–222.
- Stefanowitsch, A. 2003. Constructional semantics as a limit to grammatical alternation: the two genitives of English. In G. Rohdenburg & B. Mondorf (eds.), *Determinants of grammatical variation in English*, 413–44. Berlin and New York: Mouton de Gruyter.
- Swan, M. 1995. *Practical English usage*, 2nd edition. Oxford: Oxford University Press.
- Szmrecsanyi, B. 2004. On operationalizing syntactic complexity. In G. Purnelle, C. Fairon & A. Dister (eds.), *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10–12, 2004*, 1032–9. Louvain-la-Neuve: Presses universitaires de Louvain.
- Szmrecsanyi, B. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin and New York: Mouton de Gruyter.
- Tagliamonte, S. & J. Smith. 2002. ‘Either it isn’t or it’s not’: NEG/AUX contraction in British dialects. *English World Wide* **23**: 251–81.
- Tannen, D. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.
- Taylor, J. 1989. Possessive genitives in English. *Linguistics* **27**: 663–86.
- Thomas, R. 1931. Syntactical processes involved in the development of the adnominal periphrastic genitive in the English language. MS. Ann Arbor: University of Michigan.
- Wasow, T. 1997. Remarks on grammatical weight. *Language Variation and Change* **9**: 81–105.
- Wasow, T. 2002. *Postverbal behavior*. Stanford, CA: CSLI Publications.
- Weiner, J. & W. Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* **19**: 29–58.
- Wierzbicka, A. 1998. The semantics of English causative constructions in a universal-typological perspective. In M. Tomasello (ed.) *The new psychology of language. Cognitive and functional approaches to language structure*, 113–53. Mahwah, NJ: Lawrence Erlbaum Associates.
- Zaenen, A., J. Carlette, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M. C. O’Connor & T. Wasow. 2004. Animacy encoding in English: Why and how. In D. Byron & B. Webber (eds.) *Proceedings of the 2004 ACL workshop on discourse annotation, Barcelona, July 2004*. 118–25.
- Zwicky, A. 1987. Suppressing the Zs. *Journal of Linguistics* **23**: 133–48.