Which-hunting and the Standard English relative clause

Lars Hinrichs, Benedikt Szmrecsanyi, Axel Bohmann

Language, Volume 91, Number 4, December 2015, pp. 806-836 (Article)

Published by Linguistic Society of America

WHICH-HUNTING AND THE STANDARD ENGLISH RELATIVE CLAUSE

Lars Hinrichs
The University of Texas at Austin

Benedikt Szmrecsanyi
KU Leuven

Axel Bohmann
The University of Texas at Austin

Alternation among restrictive relativizers in written Standard English is undergoing a massive shift from *which* to *that*. In corpora of written-edited-published British and American English covering the period from 1961–1992, American English spearheads this change. We study 16,868 restrictive relative clauses with inanimate antecedents from the Brown quartet of corpora. Predictors include additional areas of variation regulated by prescriptivism. We show that: (i) relativizer deletion follows different constraints from the selection of either *that* or *which*; (ii) this change is a case of institutionally backed colloquialization-*cum*-Americanization; and (iii) uptake of the precept correlates with avoidance of the passive voice at the text level but not with other prescriptive rules.*

Keywords: restrictive relative clauses, relativizer omission, prescriptivism, logistic regression, Standard English

1. Introduction. The extent to which prescriptivist recommendations on linguistic choices can have some degree of influence on actual language use at the lexical and (morpho)syntactic levels is a matter of current debate. In corpus-based research on this question, authors have generally triangulated between linguistic precepts, as documented in the prescriptive literature, and observed changes over time in the frequency and conditioning of the variables that these precepts affect. This strand of research has faced the problem that a causal connection between linguistic precepts and observed usage is assumed, but cannot be proven, to exist. If observed usage showed a change in the direction of a linguistic precept, a causal relationship between precept and use was postulated.

In this spirit, Poplack and Dion (2009) conducted a multivariate study, based on an apparent-time data set spanning 119 years, of constraints on the expression of future temporal reference in spoken French and failed to obtain a match between the normative literature and actual usage patterns. Similarly, Anderwald (2011) shows that the popularity of nonstandard past-tense verb forms such as *she drunk* in spoken varieties of British English is robust in the face of norms prescribing distinct past-tense and past-participle forms. But other studies—especially those concerned with frequency shifts in written language—have, at a minimum, been unable to rule out an effect of the prescriptivist grammars. For example, in a study of the usage levels of verbs in the indicative and the subjunctive moods across six subcorpora of British English covering six different time points between 1570 and 1899, Auer (2006) finds that, from the mid- to late eighteenth century onward, relative frequency levels were shifting in the direction proposed by prescriptive grammarians—that is, in the direction of the tradition of gram-

* The first author gratefully acknowledges a research fellowship from the Fritz Thyssen Foundation (Cologne), which provided support for part of the work on this research. The second author gratefully acknowledges financial support by the Freiburg Institute for Advanced Studies (FRIAS) and an Odysseus grant by the Research Foundation Flanders (FWO, grant no. G.0C59.13N). We thank the following individuals, who have worked as research assistants on this project: Maren Holzkamp, Michael Percillier, Melanie Röthlisberger, Ulrike Schneider, Patrick Schultz, Gregory Webster, and Christoph Wolk. We benefited greatly from discussions with the following colleagues (aside from those cited): Lieselotte Anderwald, Jason Baldridge, Kyle Gorman, Daniel Ezra Johnson, Christian Mair, Spiros Moschonas, Tom Wasow, Christoph Wolk, and two anonymous referees. Their comments improved this article. The usual disclaimers apply.

matography that developed around the works of, for instance, Robert Lowth (*A short introduction to English grammar*, published 1762 and reissued forty-five times by 1800) and Lindley Murray (*English grammar*, 1795). Auer finds an overall trend of decreasing frequencies for subjunctive verbs, which is, however, stalled as a result of the influence of prescriptive grammatography. She frames her analysis as an illustration of the influence of prescriptivism on language use, but concedes the lack of any positive evidence supporting that causal link:

> Considering that we are not aware of any other intralinguistic and/or extralinguistic factors which are responsible for the development of the subjunctive form in the eighteenth century, it appears that prescriptivists did exert an influence. (Auer 2006:48)

The exact nature of the sociolinguistic mechanisms by which prescriptivism changes language, as well as the conditions under which it does and when it does not, marks a young area of research with many gaps, as several authors note (Auer 2006, Auer & González-Díaz 2005, Busse & Schröder 2006, 2010, Chapman 2010, Peters & Young 1997). Work in this area must begin with improved empirical description. To that end, this article applies quantitative methods to several cases of linguistic variation at once in the same data set, which documents recent change in written Standard English (StE). For the variables we chose, prescriptivists were propagating fairly consistent advice throughout most of the twentieth century and, for most of them, before that. They include calls for:

- avoidance of split infinitives,
- avoidance of stranded prepositions,
- the use of *shall* (as opposed to *will*) in verb phrases with future reference when the subject is in the first person, and
- the use of verbs in the active voice as opposed to the passive.

The first three rules emerged from the British usage literature and are quite old. Proscription of both split infinitives and preposition stranding was included in Lowth's 1762 grammar.[1] A rule aiming to regulate the semantic distribution of *shall* and *will* is already found in John Wallis's 1653 *Grammaticae Linguae Anglicanae* (Riley & Parker 1998:36).

The last of these rules, 'choose the active voice', is a concern that was added to the prescriptive core fairly recently: '19th-century writers on grammar and usage explained the structure and function of passives without any negative spin. But early in the 20th century we start to find minatory statements about it, over and over again', writes Pullum (2014, referencing Haussamen 1997:54). While prescriptivist derogation toward the passive is documented in as early a source as Woolley's *Handbook of composition* (Boston: D. C. Heath, 1907), Pullum (p.c.) suspects that it was the wide distribution of the US-American *Elements of style* (written by Strunk in 1918 and first published by his student E. B. White in 1959) that effectively canonized the precept for decades to come.

By including four prescriptivism-related variables as independent variables in the study alongside our dependent variable, the choice of relativizer in restrictive relative clauses (henceforth: RRCs), we are able to consult correlations between usage levels in the different variables as a way of shedding light on the exact mechanism through which prescriptivism changes usage.

[1] Neither of these rules originated with Lowth; rather, they were part of a set of rules that had been circulating in grammatographic discourse in England since the seventeenth century which based many rules on the grammar of Latin.

At the center of our interest here lies the choice between *that*, *which*, and *zero* in RRCs (as illustrated in 1a–c). In order to focus on variation between these three forms, this study considers only RRCs with INANIMATE antecedent noun phrases. Recent prescriptive literature has been recommending that only the option *that*, as in 1a, be used in RRCs. This precept is known as the *that*-rule (Bohmann & Schultz 2011). It can be seen as originating from a desire for grammatical symmetry: because nonrestrictive relative clauses can only take *which* as relativizers, it has seemed desirable to some grammarians, beginning with the Englishman Henry Fowler in 1926 (Fowler & Crystal 2009:634–38), to limit the range of choices in a similar way for restrictive clauses.

(1) a. This is the house **that** Jack built.
b. This is the house **which** Jack built.
c. This is the house Ø Jack built.

To be sure: all three options are grammatical under the rules of English (Biber et al. 1999:608–24, Huddleston & Pullum 2002:1047–57) in restrictive nonsubject relative clauses (i.e. clauses where the relativizer acts as direct object, indirect object, or object of preposition). *Zero*, as in 1c, is not permitted in StE restrictive subject relative clauses (which makes StE odd from a typological perspective; see Keenan & Comrie 1977).[2] A cursory glance at the history of the language shows that a categorical division of labor between *that* and *which* (and other WH-forms) according to the RESTRICTIVENESS of the clause was never part of English syntax—both types have always been part of English users' actual practice; only the animacy of the antecedent and the syntactic function of the relativizer have acted as near-categorical constraints on the forms' distribution (see §1.1 below). Accordingly, authors of prescriptivist publications on English grammar—until now—have been presenting their advice on RRCs as a recommendation, not as a rule. Fowler, for instance, asserts that 'if writers would agree to regard *that* as the defining relative pronoun, & *which* as the non-defining, there would be much gain both in lucidity & in ease' (Fowler & Crystal 2009:635). Similarly, Strunk and White emphasize the 'convenience to all' (1999:59) that would result from a clear distinction of relativizers according to restrictiveness. The widely used word-processing software Microsoft Word gives a warning to indicate detection of a 'possible grammatical error'[3] when *which* is used in restrictive relative contexts. In all published discourse on the choice of relativizer in RRCs, the option *zero* (as in 1c) plays a minor role.

Leech and Smith (2006) present initial corpus-linguistic findings on the changing frequencies of *that* and *which* in RRCs from the 1960s to the 1990s. Their study shows a substantive shift in frequencies in the direction of the precept: in both British English (BrE) and American English (AmE), the frequency of *that* has increased by several percentage points and the frequency of *which* has decreased. Their preliminary analysis of *zero* led Leech and Smith to conclude that whatever change is ongoing in the RRC of StE, the form *zero* is not affected by it. Furthermore, Leech and Smith's study suggests a methodological desideratum for work on restrictive relativizers: the automatic re-

[2] The notion of 'Standard' English here is complicated in such a way as to require some specification: it is WRITTEN StE that fully prohibits *zero* subject relativization, as our data, which do not contain any, confirm. Meanwhile, D'Arcy and Tagliamonte (2010) study spoken Toronto English as elicited in sociolinguistic interviews, a variety they refer to as 'standard' (p. 404) and 'relatively standard' (p. 389). Nonetheless, 3.6% of the subject relative tokens in their data set (which includes, unlike our data, *who* as relativizer and excludes *whose*, *which*, and *whom*) are cases of *zero* (sixty out of 1,675 cases; p. 391).

[3] So defined in the online help compendium for Microsoft Word, retrieved at http://office.microsoft.com/en-us/word-help/what-do-the-underlines-in-my-document-mean-HP005270413.aspx.

trieval of *zero* forms is much less straightforward than the search for *that* and *which*, which can be found through simple text searches. Therefore, due to the laboriousness of manually retrieving *zero*-RRCs, Leech and Smith confined themselves to searching for *zero*-RRCs in subsamples of the (unparsed) corpora they were studying (the 'Brown family' of corpora, also used in the present article). This procedure would limit the application of multivariate variationist statistics to a small subset of the data and incur a loss of statistical power. The present article presents a machine-learning-based method for the automatic retrieval of *zero*-RRCs (see §3.1 below), which enables us to apply multivariate statistics to all three variants of the relativizer.

The inclusion of all relevant cases of *zero*-RRCs is critical because, as argued above, *that* varies with *zero* as well as with WH-forms. In order to study a diachronic change in the frequency levels of *that*-RRCs, it is important to consider both alternatives in detail, if only to robustly confirm Leech and Smith's (2006) claim that the observed frequency increase for *that* is entirely compensated for by a corresponding drop in *which*, but stands in no relation with *zero*. This claim cannot be based on raw frequencies alone, but requires the simultaneous analysis of relevant competitors.

Given the large number of factors that are known to contribute to variation in the StE RRC, we propose that a fully accountable study of ongoing change in this area must avoid a monolithic explanation. Instead, we employ multivariate regression modeling and include a broad range of independent factors. A multivariate approach is necessary because it is the only way to control for the effects of multiple independents at once.

To illustrate: it is known that textual information density, as measured by type-token ratio, increased in written StE prose throughout the latter half of the twentieth century (Biber 2003, Leech et al. 2009); it is further known that greater information density predicts the choice of deleted forms (Hinrichs & Szmrecsanyi 2007:458, 461). These dynamics must be considered together with other known relevant factors in restrictive relativizer variation. To illustrate the importance of social and discourse factors, consider the stylistic value of *which*, the most formal restrictive relativizer. We hypothesize that the more formal text types included in our data would show a greater likelihood for the choice of *which*. It is imperative to study how the observed drop in the frequency of *which* from the 1960s to the 1990s is implemented across the different genres, in other words: genre must be included in a carefully designed study of this phenomenon—along with (real) time and a measure of information density such as type-token ratio.

The predictors we employ, presented in more detail in §3.2 below, fall into three types.

(i) Language-internal predictors, such as length of the relative clause
(ii) Language-external and stylistic predictors, including genre effects (Biber 1988)
(iii) Prescriptivism-related predictors, namely four other variables in addition to RRC relativizer choice that are also frequently the subject of prescriptivist discourse (Peters & Young 1997), such as the extent to which corpus texts exhibit stranded prepositions

The overall goal of our study is to describe and theoretically characterize the complex dynamics of factors under which the frequencies of *that*- and *which*-usage are shifting in the direction favored by present-day prescriptivists; thus we intend to theorize the effects of prescribed norms on actual StE usage and to determine if such dynamics can be operationalized within a variationist study design.

1.1. Variation in the standard english relative clause. Relativization is a well-researched area of English grammar, with studies existing both of standard vari-

eties and writing (Guy & Bayley 1995) and of vernacular speech (Tagliamonte et al. 2005, Levey 2006). A historical perspective is provided in, for example, Ball 1996.

Present-day StE provides a choice of nine different relativizers: *which*, *who*, *whose*, *whom*, *that*, *where*, *when*, *why*, and *zero*. A wide range of factors influence the choice from among these different forms. Two constraints regulating relativizer choice in StE are near-categorical: the animacy constraint and the restrictiveness constraint (Quirk 1957). The animacy constraint regulates relativization in favor of the choice of *who(m/se)* over other relativizers (particularly over *which*) when the antecedent noun phrase is animate (D'Arcy & Tagliamonte 2010). This constraint is relatively new to English syntax and has only operated categorically in written English since the nineteenth century (Nevalainen 2012). The restrictiveness constraint excludes *that* from nonrestrictive relative clauses. Formulations of this precept, to our knowledge, never address *zero*; instead, *which* is prescribed as the better alternative to *that*.

The two relativizers *which* and *that* are 'the most flexible in their use' as well as the most frequently used StE relativizers (Biber et al. 1999:611, 609). The deleted form (*zero*) is a possible replacement for both of these in a large number of cases, and it is 'moderately common' (Biber et al. 1999:609; in our data set, counting both subject and nonsubject contexts, *zero* is the second most frequent form). The syntactic function of the relativizer within the relative clause provides, in edited StE, a categorical constraint on the use of *zero*, which cannot be used in subject relative clauses.

Given our interest in the *that*-rule, we focus our analysis on uses of variation in those relativizer slots in which *that* is permitted: RRCs. We consider it along with interchangeable uses of *which* and *zero*.

Synchronic variation in present-day English is characterized by a clear stylistic difference among *that*, *which*, and *zero* as restrictive relativizers. In spoken, conversational language, *that* is the most frequently used relativizer, with *zero* ranking second in frequency and *which* third (Biber et al. 1999:610). Academic prose is most strongly different from conversation in this respect: Biber and colleagues find that *which* is the most frequently used relativizer in this register, with *that* in second place and *zero* in third (p. 611). The distributions in the fiction and news registers represent intermediate tendencies, with fiction (*that* > *which* > *zero*) resembling conversational speech a little more than news writing (*which* > *that* > *zero*) does. From these clear register-level preferences can be inferred the informal stylistic value of *that* (see also D'Arcy & Tagliamonte 2010) and the formal value of *which*. By contrast, frequencies of the bare form show much less variability across registers than the two overt forms, and so its primary stylistic value is much less clear.

Ever since the emergence of *which* as a relativizer in the Late Middle English period (Fischer 1992:296), *which* and *that* have held these converse stylistic connotations, paralleled by the differences in their relative frequencies in 'conceptually written vs. oral' (Koch & Oesterreicher 1985) texts (the latter being written sources that document speech, that is, texts that serve as proxies for speech data from days before the invention of audio recording). The most notable diachronic development (Ball 1996) is the rise of WH-forms to quantitative dominance beginning in the sixteenth century (for human antecedents) and the seventeenth century (for nonhuman antecedents). These variants encroached upon the territory of *that* and *zero* (Figure 1[4]). This rise of WH-relativizers includes *which* in RRCs.

[4] Following Tagliamonte et al. 2005, we visualize the data that Ball (1996) presents as two data series (table 17, p. 249) in one consolidated diagram (see our Fig. 1 and compare Tagliamonte et al. 2005:79).

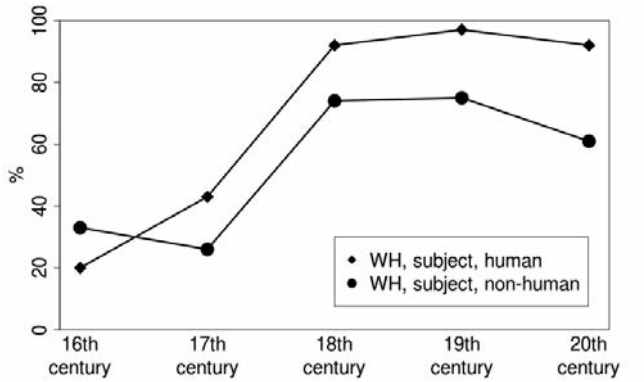


Figure 1. Wh-relativizers in subject-RRCs according to antecedent type (data from Ball 1996:249, table 17).

If one believes, as we do, that recent trends in the relative frequencies of *that*- and *which*-usage in relative clauses are in part the work of grammatical prescriptivism, then this (if you will) success appears all the more astonishing considering that it amounts to a reversal of a 500-year-old trend in relative clause formation. In her examination of historical change in the relative pronoun system (specifically, variation between Middle Scots, Modern Scots, and modern Philadelphia English), Romaine (1982:201) asks: 'Can linguistic systems change by means other than successive reweightings of linguistic features?'. The change in written StE that we are examining leads us to answer this question with a resounding 'yes': social forces such as the channels of education and publishing through which prescriptive advice is disseminated have the power to change, at the very least, discourse norms, a possible preliminary to language change.

1.2. Grammatical prescriptivism as a subject of academic discourse. As Woolard and Schieffelin (1994:69) write, 'modern linguistics has generally held that linguistic ideology and prescriptive norms have little significant—or, paradoxically, only pernicious—effect on speech forms (although they may have some less negligible effect on writing)', citing Bloomfield (1927) as an example of this position (see also Anderwald 2011, Poplack & Dion 2009). In a heightened version of this dismissal, some professional linguists take a downright hostile position toward prescriptive grammarians. A recent example of this position appears in Pullum 2009, an appraisal of Strunk and White's *Elements of style* on the occasion of the fiftieth anniversary of its first mass-market publication. Pullum's review is entitled '50 years of stupid grammar advice' (see Bohmann & Schultz 2011 for an overview of recent critiques of prescriptivism by professional descriptive linguists).

Curzan (2014:15) argues that students of language change ought to take the ideological force of prescriptivism much more seriously. While we are not aware of any empirical studies showing that prescriptivism has also influenced spoken usage, evidence of its effect on writing is available and convincing (Auer 2006, Auer & González-Díaz 2005, Busse & Schröder 2006, 2010, Cameron 1995).

Anthropological discussions of prescriptivism have pointed out that linguists' dismissive stance toward prescriptivism is problematic: it veils the fact that descriptive linguistics operates in the same paradigm of institutional enforcement as prescriptive grammar-writing does and that it is thus, to some extent, covertly complicit in the normative enterprise of prescriptivism (Woolard & Schieffelin 1994:69, Sankoff 1989).

Some recent publications take an epistemological approach to prescriptivism. Peters and Young (1997) survey forty usage guides and their treatment of eleven points of

grammar, including the choice between *that* and *which* in RRCs. They find that none of the guidebooks in its prescriptions makes reference to linguistic grammars; instead, argumentation by and large is found to be 'a strange mixture of the idiosyncratic and the conventional' (320), based in part on authors' personal aesthetic sensibilities and in part on the invocation of a core set of precepts that comprise what might be called a prescriptivist canon. Regarding the choice of relativizer in RRCs, ten of the forty usage guides state explicitly that both *that* and *which* are acceptable. Nine publications do not address the matter. While eleven of the works sampled advise against using *which*, only nine out of forty guidebooks do so categorically. The number of publications actually propagating the strong form of the *that*-rule hence is relatively small, which makes its effect on usage all the more impressive. It should be noted, however, that among the guides to support the *that*-rule are the two most widely distributed ones, the British Henry Fowler's *Modern English usage* (written in 1918 and first published in 1926) and the American *Elements of style* by Strunk and White.

Algeo (1991) provides a typology of usage guides according to the type of justifications cited for their prescriptions. His two broad classes are 'Subjective moralizing' and 'Objective reporting'. The prior refers to discourse about grammar that bases its claims about correctness and appropriateness primarily in the writer's personal tastes and delivers them with a stance of implied superiority vis-à-vis anyone who does not follow those rules. The latter reports opinions about use, actual use, or reports of actual use. Clearly, the *that*-rule, which has been part of twentieth-century prescriptive literature since the publication of the first edition of Fowler's *Modern English usage*, that is, from the very beginning of the century, was originally based in subjective moralizing and not informed by actual usage. By the end of the twentieth century, the second type of guide was able to report on usage aligning with the *that*-rule.

In her recent book-length treatment of the subject, Curzan (2014) urges linguists to take a more differentiated approach to prescriptivism in its various guises, distinguishing among four basic kinds of prescriptive rule: the 'standardizing', the 'stylistic', the 'restorative', and the 'politically responsive' (pp. 24–40). The *that*-rule falls within the second category. It does not rely on an opposition between standard and nonstandard forms (since both *that* and *which* are grammatical options in StE) and hence does not represent 'standardizing prescriptivism'. Nor does it make sense to interpret it as 'restorative' or 'politically responsive'. While the prescribed variant is indeed historically longer-established, both forms have been around for centuries and, in any case, usage guides do not mobilize historical arguments in advocating for the choice of *that* in RRCs. A political dimension, whether explicit or implicit, is likewise absent from the prescript. Relativizer choice does not reflect or index membership in any specific group but rather concerns English users across the board.

Within Curzan's typology, it may turn out that the stylistic type of rule in general is most likely to succeed in the long run. Unlike politically responsive precepts, stylistic ones do not face immediate opposition from certain groups. Nor do they explicitly run counter to the general drift of the language and its patterns of actual use, as restorative and standardizing rules tend to do. We note, however, that any given rule does not operate within a sociolinguistic vacuum, but interacts with broader developments that affect the language on a more general level. The extent to which any prescriptivist rule aligns with these general trends, we argue, plays a decisive role in determining its ultimate success.

1.3. Trends in late twentieth century standard english morphosyntax. Any short-term diachronic developments in morphosyntactic usage must be viewed

within the context of the larger trends that are known to be taking place in international StE. Whatever differences in the preferences of language users have emerged over time as a consequence of the insistence of prescriptive grammarians must be distinguished from the broad trends that have been identified in work on twentieth-century standard varieties of English (Leech et al. 2009:Ch. 11). These trends include DENSIFICATION, COLLOQUIALIZATION, and AMERICANIZATION.

Densification is the observed fact that throughout the twentieth century, information density steadily increased in written texts, especially in journalistic genres (Biber 2003). An established measure used in corpus linguistics to approximate density is type-token ratio.

Colloquialization refers to the twentieth-century tendency of written prose in English to assimilate to certain morphosyntactic and discourse features of spoken language. As Biber writes, '[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' (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' (ibid.). Most notably, these genres started displaying a dispreference for certain 'stereotypically literate features, such as passive verbs, relative clause constructions and elaborated noun phrases' (ibid.)—namely, those forms that became more frequent in the academic genres. This dissimilation of text types still continues, as Biber and Finegan (2001) have shown; in fact, it accelerated notably in the twentieth century.

Americanization refers to the observation that many broad diachronic trends, including colloquialization and densification, are more advanced in AmE than in BrE. The term *Americanization* imposes the interpretation that BrE has adopted the pattern of following the developmental trend set by AmE, an observation that is rooted primarily in the analysis of syntax (Mair 2006; for a detailed discussion of change at other levels of linguistic analysis see Hundt 2009).

1.4. RESEARCH OBJECTIVES. A corpus-based study such as the present one, concerned with changes in language use in the direction of linguistic precepts, must consider the broad trends introduced in §1.3 above and address how much of the observed change can be attributed to densification, colloquialization, or Americanization.

In order to get the most nuanced picture possible of change in the StE relative clause, a host of linguistic factors that are known to influence relativizer choice are considered. The main task for this enterprise is to cast the influence of linguistic prescriptivism on language use in greater relief in quantitative and qualitative ways. Therefore, as stated in the introduction, we consider not only language-internal as well as language-external types of factors, but also a group of predictors we call prescriptivism-related factors: we include measures of adherence to prescriptive rules as independent predictors of compliance with the *that*-rule in order to be able to gauge each writer's degree of overall adherence to prescriptive grammar. This approach will enable us (i) to diagnose, if it exists, an overall increase in adherence to prescriptive rules in written English in the latter decades of the twentieth century, and (ii) to see it in the context of other ongoing shifts in the factors conditioning variation in RRCs.

While it is by now established that *that* is encroaching on *which*'s turf in written standard BrE and AmE RRCs (Leech & Smith 2006), it has not been convincingly shown

that *zero* is not also increasingly taking the place of *which* (which would call into question prescriptivism-related accounts of the decline of *which*). This gap in the literature is due to the methodological difficulties that come with the study of *zero*-relatives: it is comparatively hard to automatically retrieve unrealized forms in a digitized corpus (see §3.1 below on the method we adopted), whereas overt forms such as *that* and *which* can be retrieved very easily, especially in part-of-speech-tagged (POS-tagged) corpora. Leech and Smith (2006) make an effort, based on a limited set of manually extracted *zero*-RRCs, to show that *zero* is not also encroaching upon the grammatical territory of *which*: frequencies of *zero* seem stable in the short-term diachronic view. Multivariate analysis is required to answer this question reliably. In keeping with the state of the art in variation research (e.g. Gries & Hilpert 2010, Johnson 2009, Tagliamonte & Baayen 2012, Wolk et al. 2013), we employ mixed-effects logistic regression modeling.

2. THE DATA. The Brown quartet of corpora is a suite of one-million-word corpora that (i) each contain 500 2,000-word samples representing different genres of written, edited, and published StE, (ii) each represent one nationally defined variety of English at a specific point in time, (iii) are of practically equal size, with minimal variance,[5] and (iv) all follow the same sampling guidelines in the selection of samples. The four corpora are as follows: Brown, comprising AmE texts from 1961; LOB (Lancaster-Oslo/Bergen), comprising BrE texts from 1961; F-LOB (Freiburg update of LOB), comprising BrE texts from 1991; and Frown (Freiburg update of Brown), comprising AmE texts from 1992; see Hinrichs et al. 2010 for more details. The list of text types sampled, along with the number of 2,000-word samples included in each category, is included in the appendix to this article (see Kučera & Francis 1967 on the compilation of Brown, Johansson & Hofland 1989 on LOB, Sand & Siemund 1992 on F-LOB, and Hinrichs et al. 2010 on the complete suite of four corpora).

3. THE LINGUISTIC VARIABLE.

3.1. THE DEPENDENT VARIABLE. Linguistic variation occurs where more than one linguistic form can be used in a syntagmatic slot, where each alternative would be equally correct according to the speech community's grammar, as well as semantically largely equivalent to the other choices (Cheshire 1987, Wolfram 1991). In other words, a study of linguistic variation focuses on the CHOICES made by speakers and writers from among several available 'alternate ways of saying "the same" thing' (Labov 1972:188). Our analysis focuses on writers' choices between *that* and its alternatives in RRCs.

Prescriptivism bears most strongly on the choice between *that* and WH-forms in RESTRICTIVE relativization. To ensure interchangeability throughout the data set we narrowed the variable context by excluding from the study RRCs that employ, or would employ, WH-relativizers other than *which*, namely *who(m/se)*. Oblique relatives with pied-piping were likewise ignored because they can take only *which* as relativizer: *that's the house about which I was talking* (but **about that/Ø I was talking*). Given these exclusions, 'complex relativizers', as in *we've never met the people whose house we are renting*, were also excluded (Huddleston & Pullum 2002:1039). We further excluded all cases of clauses in which the relativizer can be *when*, *where*, or *why* and performs an adverbial function, thus narrowing down the data set to subject- and nonsubject-RRCs.

[5] Minimal fluctuation in the overall size of the corpora results from slight differences in the length of text samples: following a general guideline, each of the 500 texts sampled for each of the four corpora will be sampled (i) at least up to the 2,000th word, but (ii) only up to the end of the sentence in which the 2,000th word occurs.

Following from the above, the primary variable we are considering is the alternation between *that* and *which* in RRCs. Since the *zero* form is also highly frequent in nonsubject-RRCs, we defined different variable contexts for subject and nonsubject-RRCs,

	that	*which*	*zero*	TOTAL
Subject-RRCs	6,312	4,818	0	11,130
Nonsubject-RRCs	1,045	1,016	3,677	5,738
TOTAL	7,357	5,834	3,677	**16,868**

TABLE 1. The data set: RRC tokens by relativizer form and syntactic function.

extracting all interchangeable cases of (i) *that*- and *which*-clauses in subject contexts and (ii) *that*-, *which*-, and *zero*-cases in nonsubject contexts. Table 1 shows the distribution of all tokens according to syntactic clause function. The total number of clauses included in this study amounts to $N = 16{,}868$.

EXTRACTION OF *which*- AND *that*-RRCS. Both of the overt relativizer forms under investigation, *which* and *that*, were easily retrieved in the Brown corpora and disambiguated from occurrences of the same forms in syntactic functions other than relativization based on the available POS tagging. All four corpora have been automatically POS-tagged in C8; in addition, F-LOB and Frown have also been manually postedited (Hinrichs et al. 2010:200). C8 applies distinct tags to *which* in relativizer function and to *which* as interrogative determiner, which enabled us to search for only tokens in the relevant grammatical function.

In order to limit clause retrieval to RESTRICTIVE relative clauses introduced by *which* and *that*, cases of relativizers preceded by a comma were excluded from the search.[6] Similarly, since the variable context excludes pied-piping, the presence of a word marked by a preposition tag directly preceding the relativizer was a reason for exclusion of the clause.

EXTRACTION OF *zero*-RRCS. While the retrieval of *that*- and *which*-relativizer tokens in POS-tagged corpora as described above is unproblematic, automatic searches for *zero*-relatives pose a considerable challenge, especially in non-grammatically parsed corpora such as the Brown corpora.[7] Linguists have often had to resort to the manual extraction of *zero*-relatives. Given the immense workload of completely extracting all cases of a single variable from linguistic corpora, authors have typically restricted their manual search for *zero*-relatives to representative subsamples of their corpora (Leech & Smith 2006, Olofsson 1981). This procedure did not seem viable for a variationist study of the full corpora of the Brown family. Using Python scripts, we initially explored a POS-pattern-searching method for *zero*-relative retrieval similar to the one described by Lehmann (2002). This approach involves defining a list of POS sequences that frequently contain *zero*-relatives, automatically retrieving all cases of these patterns from a corpus, and hand-sorting the data to weed out false hits. Much like Lehmann, we found this method to be unsatisfactory in terms of the immense numbers of both false negatives and false positives.

A great improvement in terms of recall and precision was achieved by introducing a statistical supervised machine-learning approach to the automatic detection of *zero*-

[6] Biber and colleagues (1999), Huddleston and Pullum (2002), and Quirk and colleagues (1972) agree that nonrestrictive relative clauses in present-day written StE (unlike in, say, seventeenth-century English; see Denison & Hundt 2013) are generally preceded or enclosed by comma(s).

[7] Hundt and colleagues (2012) show that retrieving *zero*-relative clauses can be difficult even in parsed material.

relatives in the POS-tagged corpora. The system we chose for the task employs a conditional random field (CRF) framework (Lafferty et al. 2001) and was trained using features defined over *zero*-relative labelings from the Penn Treebank (Marcus et al. 1993). The technical aspects of this process are described in an online appendix, which is available at http://muse.jhu.edu/journals/language/v091/91.4.hinrichs01.pdf.

MANUAL MARK-UP. Following retrieval of RRCs in the four corpora, the entire data set was hand-coded as follows:

- a standardized tag in angular brackets was inserted right before the beginning of the antecedent NP ('<$W>', '<$T>', and '<$Z>' for *which*, *that*, and *zero* respectively),
- a corresponding closing tag was inserted right after the end of the relative clause ('<%W>', '<%T>', and '<%Z>'), and
- for *that*- and *which*-RRCs, the POS-tag marking the relativizer was manually modified to reflect whether it introduces a subject- or nonsubject-RRC. For *zero*-RRCs, a tag was inserted to mark the deleted relativizer's position in the object gap of the relative clause.

The resultant codings for *that*-, *which*-, and *zero*-RRCs, respectively, are illustrated in 3.

(3) a. **<$T>** a spontaneous popular uprising **<W WPRO>that** the guerrillas could not ignore **<%T>** (F-LOB-A04)

b. **<$W>** the patrol **<DDLO>which** Sergeant Prevot led out that next night **<%W>** (Brown-K02)

c. **<$Z>** some kind of trick **<W ZR>** Budd had thought up **<%Z>** (Brown-N01)

3.2. INDEPENDENT VARIABLES. The corpora containing mark-up for all 16,868 RRCs were then further analyzed in order to annotate each token for a wide range of independent variables. This part of the analysis used a suite of Python scripts that targeted each relativizer token and extracted all of the context variables that are introduced below. These independents can be categorized as internal (or linguistic) predictors, external and stylistic predictors, and prescriptivism-related predictors. Each predictor is briefly presented below.

INTERNAL PREDICTORS.

Relativizer function: The syntactic function of the relativizer in the clause it introduces, with the levels 'subject' and 'nonsubject'. All three relativizers may appear in nonsubject position; 'subject' only allows for *that* and *which*.

RC length: Length of the actual relative clause in words, which is a measure of the complexity of the clause introduced by the relativizer. More complex clauses are hypothesized to favor overt marking for ease of information processing (cf. Rohdenburg's (1996) 'complexity principle'). Tagliamonte and colleagues (2005:101) confirm this for northern British vernacular varieties, finding that 'short clauses favor *zero* and long clauses disfavor *zero*' in both subject and nonsubject position. This factor was log-converted.[8]

[8] When including a scalar predictor in a regression model, one entertains the possibility that the log-conversion of a factor may correlate better with the observed response curve than the raw, unconverted data for that predictor. In order to determine whether it is the raw data for the predictor or its log-converted form that fits the data better, one runs a simple, unifactorial logistic regression for the dependent variable and (in R) extracts the deviance value from the model summary. Customarily, a final model will include the form of the scalar predictor that produces a LOWER deviance value. The following code was used to determine deviance for a log-converted factor **scalarPredictor** in a linear regression model for the response: summary(glm(response ~ log(scalarPredictor), data, family="binomial"))$deviance

Preceding relativizer: The relativizer used previously to the current one in the same document. This includes the levels 'that', 'which', 'zero', and 'none'; the latter applies when the relativizer in question is the first one encountered in a document. This variable gauges structural persistence (Szmrecsanyi 2006).

Nested: A binary measure (levels: 'yes' and 'no') of whether or not the relative clause at hand is nested in another relative clause. For the genitive alternation in written StE, Hinrichs and Szmrecsanyi (2007) find a significant tendency for nested constructions to display a *horror aequi* effect, that is, a tendency to alternate from *s*-genitive to *of*-genitive and vice versa, rather than employing either variant twice in a row. We expect a similar tendency for relativizer choice in nested relative constructions.

Antecedent POS: Part of speech of the antecedent head. The range of possible parts of speech was simplified to a binary distinction between 'noun' and 'other' (e.g. pronouns, numerals, etc.). Tagliamonte and colleagues (2005) model 'type of antecedent' as a predictor. However, their factor collapses a distinction between pronouns and NPs with the question of definiteness, as does Tottie and Harvie's (2000:214) '[g]rammatical category of the antecedent NP head'. In both cases, the factor produces ambiguous results across the varieties studied. By modeling the POS of the antecedent head as a separate variable from its definiteness (see below), we hope to achieve a clearer analysis of both of these predictors.

Antecedent number: Grammatical number of the antecedent head. This is an automated measure that extracts information from the antecedent head's POS tag. For some items, grammatical number is not marked in the Brown corpora (e.g. *all* is marked '<DB>', without an identifying '1' or '2' for number). In these cases, the script defaults to the level '0'—for ambiguous—in addition to '1' for singular and '2' for plural. Tottie and Harvie (2000:208) mention quantification of the antecedent as a potential coding factor, but they neither include it in their own study nor make reference to any others who do. Rickford's (2011) extensive comparative study of relativizers in multiple spoken varieties of English coded for 'plurality of the antecedent', but did not find a significant effect for the predictor.

Antecedent length: Length of the antecedent noun phrase in words. This variable reflects the complexity of the noun phrase modified by the relative clause in question. While the surface complexity of the relative clause has been found to exert a significant effect on relativizer choice (see above), to our knowledge the same has not been tested for the antecedent noun phrase. Here, processing constraints have only been modeled in terms of adjacency of antecedent head and relative clause (Guy & Bayley 1995:154–55), which disregards the potential effects of, for instance, complex premodifiers. Again, Rohdenburg's (1996) complexity principle would predict usage of explicit relativizers in complex environments.

Definiteness: Definiteness of the antecedent head. Distinguishes overtly definite NPs (those preceded by a definite article, a demonstrative or possessive pronoun, or genitive '-s') from indefinite ones. This is a binary variable with the levels 'def' and 'indef'. Tagliamonte and colleagues (2005:100) find indefinite antecedent NPs to favor *zero* compared to definite ones.

Head-to-relativizer distance: The number of words between the antecedent head and the relativizer. In most cases the relative clause follows its antecedent head immediately, but sometimes other postmodifiers intervene, such as prepositional phrases. Guy and Bayley (1995:154–55) claim that such intervening material favors overt relativizers. However, the same factor fails to reach significance in Tagliamonte et al. 2005. We attempt to resolve these conflicting results by supplanting the binary distinction between adjacent and nonadjacent with a more detailed, continuous measure of distance.

External and stylistic predictors.

Category: Text category of the sample in which the relativizer occurs (see appendix).

Genre: Text genre of the sample in which the relativizer occurs. This variable captures the four genre-metacategories in the Brown corpora with the levels 'press', 'generalprose', 'learned', and 'fiction' (see the appendix on the grouping of the fifteen different text categories into the four genre groups).

TTR: Type-token ratio. Calculated by dividing the number of unique word types by the number of individual words (tokens) in the text. A proxy for information density, it is true that TTR is not necessarily a perfect measure, but it is one that is customarily used in variation studies (for example, it is one of the features drawn upon in Biber 1988 and follow-up research). Higher TTR values generally predict the selection of shorter forms, for instance, the choice of the *s*-genitive over the *of*-genitive in the Brown corpora (Hinrichs & Szmrecsanyi 2007). This factor was log-converted. Note that TTR does not remain constant across text sizes (Tweedie & Baayen 1998). While the standardized sampling size of the Brown corpora mitigates this problem, a word of caution is in place against extrapolating from the present study's findings to studies in which (sub)corpus size is more variable.

Mean word length: The average length in letters of a word in the given text sample. This variable reflects the lexical complexity of the corpus text under analysis. It is one of several measures of textual complexity, along with **mean sentence length** and **subordinating conjunctions**.

Mean sentence length: The average length in words of a sentence in the given text sample. This variable approximates the syntactic complexity of the corpus text under analysis. Along with **mean word length** and **subordinating conjunctions**, this variable gauges textual complexity. This factor was log-converted.

Subordinating conjunctions: Relative frequency of subordinating conjunctions, normalized to a value per 10,000 words in the corpus text under analysis. This statistic is the third measure of textual complexity of the text as a whole, in addition to **mean word length** and **mean sentence length**.

Nouniness: Relative frequency of nouns in the text sample under discussion, normalized to a value per 10,000 words. This statistic has been discussed as expressive of information density, with increasing frequencies of nouns showing one side of an overall trend of the 'densification' of written StE (Leech et al. 2009:211).

Noun-verb ratio: Ratio between number of noun-tagged and verb-tagged words in the text sample: an alternate measure for the phenomenon operationalized by **nouniness**.

Personal pronouns: Relative frequency of personal pronouns in the corpus text under analysis, normalized to a value per 10,000 words. The use of personal pronouns can be taken as an indicator of involved style (Biber 1988).

Prescriptivism-related predictors.

Stranding: Proportion of stranded prepositions out of all prepositions in a given corpus text, multiplied by 100 to yield a percentage. Some style guides explicitly proscribe 'ending a sentence with a preposition' (*the house which he looked at*), although the trend in the recent usage literature is toward a more descriptive perspective (Busse & Schröder 2006:464–66, Huddleston & Pullum 2002:138).

Split infinitives: Relative frequency of split infinitives (as in *to boldly go*) in a given corpus text, normalized to a value per 10,000 words. Split infinitives continue to draw negative commentary in many usage manuals (Busse & Schröder 2006:465–68).

Passives: Fraction of passive constructions over active lexical verbs in a given corpus text. The passive voice (as in *the motion was tabled*) is rejected outright by lan-

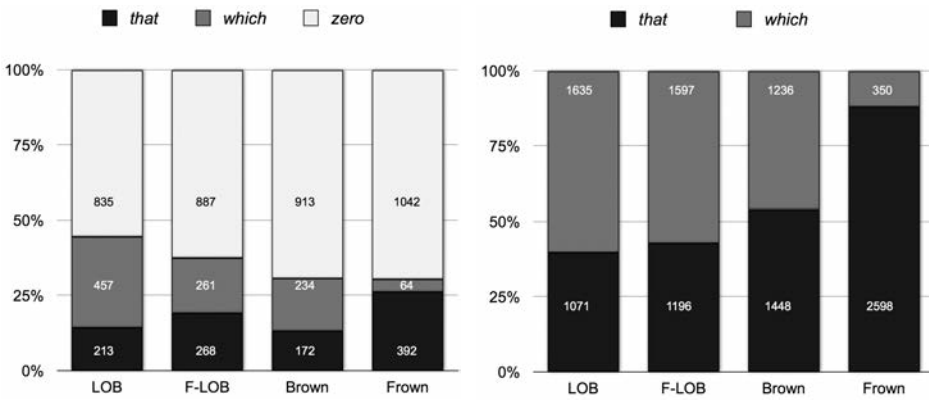
guage mavens, many of whom associate it with vagueness (Strunk & White 1999: 18–19). Leech and Smith (2009) trace a decline of the passive voice in written English during the twentieth century, which they attribute to 'prescriptive forces' (p. 183). This factor was log-converted.

Shall-will ratio: Ratio between tokens of modal verbs *will* and *shall*. Some usage-guide writers (e.g. Strunk & White 1999:58) recommend *shall* as a future marker for first-person subjects (see Facchinetti 2000 for a historical perspective). Due to the affordances of the POS tagset applied to the data, extraction of this factor was not sensitive to person.

3.3. Distribution of *which*, *that*, and *zero* across corpora. Below we show two different ways of slicing the data. First, we focus on the subset in which there is a three-way alternation between *that*, *which*, and *zero*, that is, nonsubject-RRCs, and their distribution across the four corpora (Figure 2a). The visualization clearly shows the increase of *that* both from LOB to F-LOB (i.e. on the British axis) and from Brown to Frown (American). It is notable that the change is much more pronounced in AmE: in the 1960s, AmE (Brown) shows fewer cases of *that* than BrE does (LOB); in the 1990s, that relationship is reversed. To complement Fig. 2a, Figure 2b shows the distribution of relativizers in subject-RRCs: necessarily, only *that* and *which* participate here, as per the rules of StE which, in our data, are enforced by professional editors, and which prohibit *zero* subject relativizers.

Frequencies of *zero* are proportionally stable over time in the AmE data: in Brown, 69.22% of nonsubject-RRCs are *zero*-clauses (N = 913), and while the absolute frequency rises to N = 1,042 in Frown, *zero*-clauses still account for practically the same share of the total at 69.56%. The fact that there is an overall greater number of relative clauses in AmE in the 1990s than in the 1960s points toward a possible general development toward syntactically more complex writing. These discourse-level changes are addressed in the design of our multivariate approach.

Meanwhile, *zero*-clauses in BrE writing increase both in relative and in absolute numbers. As the multivariate analysis presented in §4 below shows, this change is not found to correlate with writers' uptake of the *that*-rule or other prescriptive rules.



a. Distribution of *which*, *that*, and *zero* in nonsubject-RRCs (N = 5,738).

b. Distribution of *that* and *which* in subject-RRCs (N = 11,131).

Figure 2. Distribution of variables across corpora. Box sizes show proportions; numerical labels show raw *N*s.

Next, we re-sort the data to include both major syntactic types of RRCs, subject and nonsubject, but only *that*- and *which*-clauses (Figure 3). This visualization shows even

more clearly the reciprocal diachronic relation between the two overt forms of the relativizer. In addition, the lead of AmE in the shift becomes clearer in this more global view: both varieties shift toward using *which* less and *that* more frequently. But even by the 1990s, BrE has not quite attained the overall proportion of *that* usage that AmE already had in the 1960s.

The multivariate conditionings of the distributions shown here are considered in more detail below.

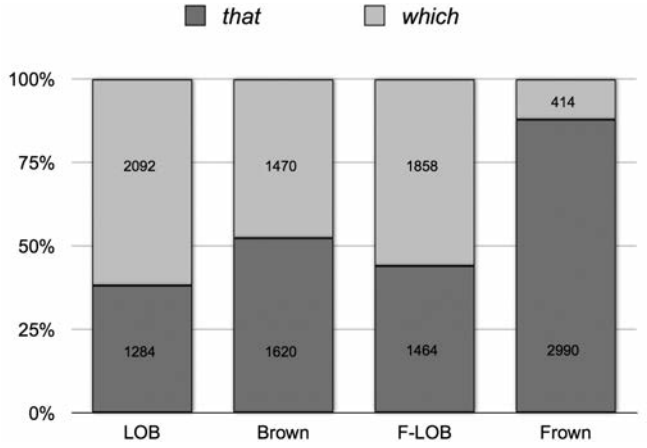


FIGURE 3. Variation between *that* and *which* in both subject- and nonsubject-RRCs ($N = 13{,}192$).

4. RELATIVIZER DELETION.

4.1. MODEL BUILDING. The multivariate procedure employed in §4 and §5 is mixed-effects logistic regression.[9] In keeping with previous work on variation between *which*, *that*, and *zero* in English relative clauses (e.g. D'Arcy & Tagliamonte 2010, Guy & Bayley 1995), regression models are fitted to binary responses, even when a case of ternary variation (i.e. variation between three options) is considered: it is modeled as the choice between one of the three options or either (or both) of its two alternatives. At no cost of statistical accountability, the dynamic reduction of complexity from a ternary to a binary response provides a considerable gain in clarity. This section considers alternation between *zero* as one option and either *that* or *which* as the other. Because *zero* does not occur in subject-RRCs, the data set for this analysis was restricted to nonsubject-RRCs. In order to ensure interchangeability among all values of the dependent, only nonsubject-RRCs with inanimate antecedent NP heads were considered ($N = 5{,}738$; see Table 1 and Fig. 2a above).[10] The selection of independent variables for inclusion in an optimal model to explain variation in the dependent variable is the central task in model fitting. We followed the customary steps in constructing the minimal adequate model. In a recent discussion of the procedure in the context of linguistic data, Gorman and

[9] Here and in §6, we utilized the implementation of generalized linear mixed-effects models in the lme4 library (R package version 0.999999-2) in R (version 3.0.1; R Development Core Team 2013).

[10] Ours is the first variationist study of English relativization that restricts itself to inanimate antecedents. Compare, for example, D'Arcy and Tagliamonte (2010), who, accepting categorical distribution of the form *who*, include both animate and inanimate antecedents in their study. While inanimates emerge, predictably, as excluded from relativization with *who* (p. 392), relativization of animate antecedents is shown to be a site of rich, socially conditioned variation in which speaker gender, interviewer gender, and (most robustly) the dyadic combination of the two can be shown to correlate with relativizer choice.

Johnson (2013:221) write that an initial model should include 'any predictors the experimenter has recorded and thinks might influence the outcome'.[11] The analyst should assess each independent variable in the resulting model based on its significance, sign, and relation to the research hypothesis, and then decide whether to leave the factor in during subsequent iterations of the model (Gorman & Johnson 2013:222).

Due to the large number of independents in our design, we did not use the customary significance threshold of $p = 0.05$ for inclusion of a factor in the model. In order to produce models with manageable numbers of degrees of freedom, we lowered the significance threshold to $p = 0.01$. The overall significance of factors was determined using the type II Wald chi-square tests performed using the Anova() utility function that is part of the car package in R (Weisberg & Fox 2011). Note, however, that the probabilities we report are taken from the model output from the lme4 package (Bates et al. 2011), which reports type II probabilities as well, but breaks them down to individual factor levels. This leads to the appearance of a more conservative treatment of factor probability. Multicollinearity is not an issue, as the model's condition number ($\kappa = 7.7$) is well below the customary threshold of 15, which indicates medium collinearity.[12] All results reported below as significant are also stable under bootstrap validation.[13]

Table 2 shows the full range of factors that were considered in the model-building process. Of the four factors relating to prescriptive recommendations presented in §3.2 above—**passives**, **shall-will ratio**, **split infinitives**, and **stranding**—only two could be shown to correlate significantly with any of the outcomes studied in §4 and §5: passives and stranding contribute to the explanation of variation between *that* and *which*, discussed in §5 below; but as Table 2 shows, none of them reaches significance for the prediction of the choice of *zero* in nonsubject-RRCs.

A range of interaction terms were tested based on the hypothesis that the external factors time, genre, and variety interact with the prescriptive factors. Also, a three-factor interaction **time** × **variety** × **genre** was tested. The inclusion of an interaction term always also implies the inclusion of each participating factor as a main effect. Main effects that entered the model in this way as part of a significant interaction term were kept in the model even if the main effect itself was not significant.

Two random effects were included. The first of these, **1 | file**, captures individual bias at the level of the corpus text sample; it thus approximates a by-subject random effect that is by now customarily included in the multivariate study of language variation. The second random effect, **1 + corpus | category**, is designed to capture any variation at the level of the fifteen different text categories that form the design of the Brown corpora. In order to make the effect sensitive to differences between the four corpora in random variation among categories, this effect fits a random intercept for each level of **category** and random slopes for **corpus**.[14]

[11] Gorman and Johnson's (2013) presentation is an adaptation of the four steps suggested by Gelman and Hill (2007:69).

[12] The condition number was calculated using R function collin.fnc(), library languageR (Baayen 2008:182).

[13] Here and in §6, we followed the bootstrap procedure suggested in Baayen 2008:283: 100 runs, sampling with replacement. The confidence intervals did not include 0.

[14] The first random effect is essentially a control for subject bias, because each file is written by a different author. The second random effect includes a random intercept for each category and a random intercept for the category-by-corpus interaction. Random effect 2 thus targets bias at the level of genre (or discourse); it is needed as a supplement to the subject random effect. Nesting is not a problem: all four corpora have the same genre structure, but file names are unique. They are specified as a combination of corpus and file ID, for instance Frown_A01, Frown_A02, and so on.

4.2. FINDINGS. The minimal adequate model for the deletion of the relativizer in nonsubject-RRCs is presented in Table 2; factors that could not be shown to make a significant contribution to the prediction of the choice of *zero* (i.e. relativizer deletion) in the nonsubject-RRC data set are excluded from the model and not shown in Table 2. The model correctly predicts 78.7% of all relativization outcomes in the data set and comes with an index of concordance (*C*) value of 0.87, indicating that the model discriminates well between *zero* and *that*/*which*.

It is most immediately striking that the choice of *zero* is in no significant statistical relationship (as main effect or interaction term) with any of the predictors relating to prescriptive recommendations. That is to say, the uptake of the two classic prescriptivist rules 'avoid the passive voice' and 'do not use stranded prepositions' correlates in no systematic way whatsoever with either a preference or a dispreference for the deletion of relativizers. The insignificance of all interaction terms involving either **passives** or **stranding** further supports the finding that the choice of *zero* is simply not affected by prescriptivism, at least as represented by the four canonical precepts we tested for.

		ODDS RATIO	*b*	*p*	
(model intercept)		1.08	0.081	0.547	
INTERNAL PREDICTORS					
preceding relativizer	*that* (default: none)	0.78	−0.249	0.041	*
	which (default: none)	0.66	−0.413	0.001	**
	zero (default: none)	1.09	0.082	0.525	
antecedent length		0.74	−0.307	0.000	***
RC length		0.21	−1.546	0.000	***
EXTERNAL AND STYLISTIC PREDICTORS					
variety	AmE (default: BrE)	1.66	0.508	0.000	***
TTR		4.21	1.439	0.000	***
subordinating conjunctions		0.99	−0.003	0.000	***
personal pronouns		1.01	0.001	0.000	***
PRESCRIPTIVISM-RELATED PREDICTORS					
—			—		
RANDOM EFFECTS					
1 | file		intercept, $N = 1{,}700$, variance: 0.4884			
1 + corpus | category		intercept, $N = 15$, variance: 0.1285 (intercept), 0.0172 (F-LOB), 0.0537 (Frown), 0.0757 (LOB)			
SUMMARY STATISTICS	*N*	5,738			
	correctly predicted	78.7% (baseline: 64.1%)			
	Somers's D_{xy}	0.73			
	C	0.87			

TABLE 2. Minimal adequate logistic mixed-effects regression model for variation between *zero* and *that*/*which* in nonsubject-RRCs (model 1). The predicted value is *zero*.

As for the language-internal predictors that survived the model-fitting stage, we note that **preceding relativizer**, that is, the choice of relativizer in the preceding slot, makes an altogether highly significant contribution to the prediction accuracy of model 1. The model shows that both *that* and *which* in the previous slot make a negative prediction for the choice of *zero*, because PERSISTENCE of the previously chosen form is most likely. The choice of *zero* in the previous slot is rated with a positive coefficient, although this factor level's *p*-value does not reach significance according to our strict alpha-value. In other words, *zero* does not prime well, at least not as well as the overt

forms do, a finding that is not unexpected given the literature on lexical enhancement effects in syntactic priming (see e.g. Gries 2005, Pickering & Branigan 1998, among others). Two more predictors complete the group of significant language-internal factors: **antecedent length** and **RC length**. For each word that is added to the length of the antecedent NP, the odds of the relativizer being deleted decrease by a factor of 0.74. For each one-unit increase in the log length of the RRC, the likelihood of *zero* being chosen decreases even more drastically, by a factor of 0.21. The predictions made by these factors are both rooted in the fact that longer phrases and clauses increase processing complexity, and the processing of complex linguistic material is aided more by the presence of overt than by deleted grammatical forms (see e.g. Jaeger 2006:102–3).

How do the language-external and stylistic predictors fare? The choice of *zero* is remarkably insensitive to variation in the external predictors: both **time** and **genre** fail to reach statistical significance and are therefore absent from the model. We do note, however, a significance of **variety** in the deletion of relativizers. The overall distribution of relativizers in nonsubject-RRCs (Fig. 2a) showed that the two American corpora, Brown and Frown, have the highest frequencies of *zero*. And in fact, the multivariate analysis shows that **variety** has a significant effect on the selection of *zero*: as per the odds ratio, a nonsubject-RRC written by an American writer is 1.66 times as likely to lack an overt relativizer as one written by a British writer.

	BrE		AmE	
	LOB	F-LOB	Brown	Frown
TTR	0.391	0.422	0.421	0.423
log(TTR)	−0.052	0.022	0.019	0.026

Table 3. Mean TTR and log-conversion of centered TTR by corpus.

Second, we observed that the frequency of *zero*-RRCs increased from the 1960s to the 1990s both in BrE (by 5.2%) and in AmE (by 11.4%). Contrary to expectation, however, the passing of **time** per se cannot be shown to have a significant effect on relativizer deletion. Instead, a look at the aggregate file statistics shows that the increase is explained by an increase in information density, measured by the factor **TTR**, which robustly predicts the choice of *zero*: in texts with higher information density, writers are more likely to choose the more economical variant. (Note that this tendency has to do with a text's overall information density, as reflected in its type-token ratio: incidences of local grammatical complexity still favor the choice of overt forms over *zero*.) The dynamic of greater textual information density favoring *zero* has remained broadly the same between the 1960s and 1990s in both varieties. But since information density has been steadily increasing in published prose since the middle of the twentieth century (as Biber has shown, e.g. Biber 2003 for newspaper discourse), the choice of economical variants has become more frequent overall. Hinrichs and Szmrecsanyi (2007) show that the same dynamic accounts for an increase in the *s*-genitive (in text categories A and B of the same corpora), which is a more economical choice than its alternative, the *of*-genitive. As Table 3 confirms, the mean information density measure of **TTR** has increased notably between the 1960s and 1990s both at the level of outright type-token ratio and (more clearly) in the log-converted measure, which entered the multivariate calculation (the log-converted factor was centered around *zero* using overall mean before entering the model statistics). While information density as expressed by **TTR** thus displays a robust and significant probability in favor of deletion, the factor **subordinating conjunctions**, which operationalizes syntactic complexity in the textual environment of the variable context, predicts nondeletion. The result confirms

that as writing style becomes more complex, with writers employing more hypotactic connections that use subordinating conjunctions, they are less likely to delete relativizers. In other words, overt relativizers tend to be retained in more complex textual structures, because, arguably, they enhance processability. This tendency is consistent across all corpora and rated as highly significant at $p \cong 0.000$. Finally, the factor **personal pronouns** operationalizes the frequency of personal pronouns in the text sample surrounding each RRC token. Higher frequencies of personal pronouns are generally considered to be indicative of more informal and involved, less informational writing style (Finegan & Biber 1994). The fact that the sign of the coefficient (*b*) for this factor is positive indicates that *zero* is the more informal option (Biber et al. 1999).

In summary, relativizer deletion in nonsubject-RRCs is predicted by factors that mostly relate to language processing, where the aspects of *zero* as an economical variant on the one hand and the property of overt relativizers as means of creating clarity and mitigating complexity (in the sense of Rohdenburg 1996) on the other are important. In addition, the stylistic value of *zero* as a marker of informality was shown to be relevant. We find no evidence that the probabilistic grammar underlying relativizer deletion in written StE has changed in any important way from the 1960s to the 1990s; the most important change relating to relativizer deletion in the data at hand is an increase in information density, which confirms the findings of publications such as Hinrichs & Szmrecsanyi 2007, Szmrecsanyi & Hinrichs 2008, and Biber 2003. Most crucial to the research question at hand is the complete lack of evidence for any statistical correlation between relativizer deletion and the uptake of any of the prescriptivist recommendations that we tested for: the choice of the active over the passive voice, the avoidance of preposition stranding, the use of *shall* in future VPs, and the avoidance of split infinitives.

5. Variation between *that* and *which*. Section 4 has described the variation between overt and deleted relativizers in nonsubject-RRCs as fundamentally constrained by factors relating to language processing, textual organization, and stylistics. We now explore the possibility that variation among *that*, *which*, and *zero* is NOT an actual case of three-way alternation. Rather, language users may make a primary choice between an overt and a deleted relativizer, a choice that is not influenced by the prescriptivism-related predictors in our study. Once an overt relativizer is chosen, a selection between *that* and *which* is made. Based on these considerations, we now abandon the issue of deletion and fit another regression model to the binary choice between *that* and *which*.

Effectively slicing out a variant from the analysis and thus reducing a three-way alternation to a two-way alternation is admittedly not entirely conventional, particularly not in work in the variationist sociolinguistics tradition, so a few justificatory remarks are in order here. The present study broadly follows the variationist methodology, in which Labov's 'foundational' (Tagliamonte 2012:9) principle of accountability takes center stage: 'any variable form … should be reported with the proportion of cases in which the form did occur in the relevant environment, compared to the total number of cases in which it might have occurred' (Labov 1969:738, n. 20). We take pains to report these proportions (see Figs. 2 and 3, as well as model 1 reported in §4), thus making sure that our analysis is fully accountable and replicable. As far as the regression model reported in this section (model 2) is concerned, we take the liberty of restricting attention to the variation between the overt markers *which* and *that*, ignoring, for the sake of simplicity, the *zero* variant for the moment. In any event, we hasten to add that it is not as though we are slicing out a huge section of the data: *zero* is outnumbered by a ratio of about four to one by the overt variants in the full data set (see Table 1).

A second way in which the model reported in this section departs from much previous work on relativizer variation is that it jointly analyzes subject AND nonsubject contexts. This is legitimate, since this article is primarily concerned with variation between *that* and other relativizers, not with differences between subject- and nonsubject-RRC contexts. Notice, however, that we do include the predictor **relativizer function**—which distinguishes between subject- and nonsubject-RRC contexts—in the regression model, for the sake of doing justice to variational differences between the two contexts.[15]

We define our data set accordingly: all *zero*-RRCs are excluded. All cases of *that*- and *which*-RRCs, with relativizers in both subject and nonsubject function, are included, and the syntactic function of the relativizer is entered as a binary factor in the analysis. The total *N* is 13,192.

5.1. Model building and findings. Model building for this analysis follows the principles laid out in §4.2: the same range of factors (with the addition of **relativizer function**) were tested using the same process of stepwise iterative modeling. The model's condition number κ is 17.6, which is another way of saying that the model exhibits moderate, though not harmful (Baayen 2008:182), collinearity. All results reported as significant below are also stable under bootstrap validation. The minimal adequate mixed-effects logistic regression model, which is shown in Table 4, correctly predicts 83.4% of all relativization outcomes in the data set and comes with an index of concordance (*C*) value of 0.92, indicating that the model discriminates very well between *that* and *which*. As for the fixed effects in the model, two observations stand out:

- The factor **time** is significant both as a main effect and as part of two different interaction terms (with **variety** and with **stranding**). In other words: unlike in model 1, in model 2 the constraints underlying the choice between *that* and *which* appear to have undergone change between the 1960s and the 1990s.
- Of our prescriptivism-related predictors, two emerge as significant: **passives** and **stranding**; in addition, **stranding** interacts significantly with **time**, suggesting not only that prescriptivism plays a role in the variation between *that* and *which*, but also that there has been a change in its role between the 1960s and the 1990s (all three terms are highly significant at $p \cong 0.000$).

The first of the three language-internal factors in the model, **preceding relativizer**, unsurprisingly makes a significant prediction: when *that* was selected in the most recent RRC, the odds of a repeated choice of *that* increase by a factor of 1.28 (compared to the default condition 'none' for cases where no previous choice can be specified because they are the first RRC in the text sample). By contrast, when the preceding choice of restrictive relativizer has been *which*, the odds of *that* decrease by 30%. As in model 1, the choice of *zero* in the previous RRC is not rated as making a significant prediction. Moving on in Table 4, we observe that subject contexts (**relativizer function**) favor *that*, and so do antecedent heads other than nouns (**antecedent POS**), which is another way of saying that nonsubject contexts and nominal antecedents favor *which*. The other two language-internal factors in the model, **antecedent length** and **RC length**, both specify, by their negative coefficients, that for each word that is added to the length of the antecedent NP or for each one-unit increase in the log length of the relative clause,

[15] The predictor **relativizer function** is specifically modeled as a main effect. We spot-checked interaction effects between **relativizer function** and the other language-internal predictors in the model, and found that against the backdrop of our research questions these lacked sufficient explanatory power to warrant inclusion in the model.

the choice of *that* becomes less likely. This observation makes an important point about the relation between *that*/*which* and processing. As antecedents and relative clauses become longer, and processing load becomes larger, writers increasingly favor overt forms over *zero*, as model 1 showed—and among the overt relativizers, they favor *which*. This shows that among the three relativizers in this study, *which* is the one that is most trusted with the task of cognitive organization and disambiguation. Thus, relative to the alternative overt form *that*, *which* seems to facilitate processing in cognitively complex environments.

With regard to language-external predictors, the fact that both **variety** and **time** make significant predictions—with AmE and the 1990s predicting considerably higher odds in favor of the choice of *that* over *which*—shows that there is a substantial portion of the shift from *which* to *that* which can be captured only as a function of intervarietal variation and variation over time. As for **genre**, there is a tendency that, vis-à-vis press texts, fictional texts favor *that* while general prose and learned prose favor *which*.

THE UPTAKE OF PRESCRIPTIVIST RULES ACROSS TIME AND SPACE. While a writer's propensity to split infinitives or to use *shall* rather than *will* in future verb constructions (with first-person subjects) correlates in no way at all with statistical preference for *that* in RRCs, as expressed by the insignificance of **split infinitives** and **shall-will ratio**, the two other prescriptivism-related factors make very clear and significant predictions. As for *shall* and *will*, we submit that given the extreme infrequency of *shall* in twentieth-century English, variability between *shall* and *will* is itself too small to matter to an area of grammatical variation in which variability is so muscular, like restrictive relativizer choice.

The factor **passives** works in the expected direction: the coefficient is negative, indicating that as the number of passive constructions relative to active constructions increase within a text sample, the choice of *that* becomes less likely. This is another way of saying that writers who disobey the prescriptive rule 'avoid passives' are also likely to ignore the '*that*-rule' for RRCs. This factor is highly significant. As Mair and Leech (2006:331–32, 337) show, the frequency of passive-voice verbs has decreased over time in the Brown family of corpora, a fact they attribute to the influence of prescriptivism. The statistical correlation we find between lower frequencies of passive verbs at the text level and adherence to the *that*-rule thus runs parallel to a similarity between the active-voice rule and the *that*-rule in the prescriptive literature. Note, however, that we do not detect an interaction between the predictors **passives** and **time**, which indicates that the active-voice rule and the *that*-rule seem to have had a similar status in the practice of writers and editors throughout the three-decade period covered by the corpora.

The second prescriptivism-related factor, **stranding**, which likewise makes a highly significant prediction, does NOT show observance of the *that*-rule and of the anti-stranding rule to be positively correlated: overall, those text samples that have higher percentages of prepositions in stranded position, that is, in clause-final position, favor *that* as restrictive relativizer and disfavor *which*.[16]

[16] A referee points out that writers with an increased preference for *that* might, therefore, also disprefer pied-piping, which requires *which* and can be seen as an alternative to preposition stranding. (The implication is one of possible collinearity for structural reasons between the **stranding** factor and *that* as dependent variable.) We remind the reader that our statistics for stranding are more global and do not study stranding as a case of variation. Furthermore, (i) our statistics quantify stranding, not pied-piping; (ii) we count stranding in any syntactic context, including nonrestrictive relative clauses, and not just in the clearly delineated contexts in which our variable occurs; and (iii) while pied-piping is one alternative to stranding, there are also others; for example, stranding in *this is the house that I closely looked at* can be avoided by lexical substitution, as in *this is the house that I examined*, and so on.

		odds ratio	*b*	*p*	
(model intercept)		0.55	−0.604	0.001	**
INTERNAL PREDICTORS					
preceding relativizer	*that* (default: none)	1.28	0.245	0.003	**
	which (default: none)	0.70	−0.354	0.000	***
	zero (default: none)	0.90	−0.101	0.280	
relativizer function	subject (default: nonsubject)	1.42	0.351	0.000	***
antecedent POS	other (default: noun)	1.68	0.518	0.000	***
antecedent length		0.95	−0.053	0.000	***
RC length		0.51	−0.675	0.000	***
EXTERNAL AND STYLISTIC PREDICTORS					
time	1990s (default: 1960s)	1.38	0.324	0.005	**
variety	AmE (default: BrE)	8.41	2.130	0.000	***
genre	fiction (default: press)	1.52	0.416	0.057	.
	general prose (default: press)	0.68	−0.389	0.028	*
	learned (default: press)	0.61	−0.490	0.026	*
PRESCRIPTIVISM-RELATED PREDICTORS					
passives		0.67	−0.401	0.000	***
stranding		1.54	0.434	0.000	***
INTERACTION EFFECTS					
variety × genre	AmE : general prose	0.31	−1.177	0.003	**
	AmE : learned	0.29	−1.252	0.005	**
	AmE : fiction	0.19	−1.660	0.000	***
time × variety	1990s : AmE	7.21	1.975	0.000	***
stranding × time	1990s	0.66	−0.420	0.000	***
RANDOM EFFECTS					
1 | file		intercept, $N = 1{,}944$, variance: 2.0257			
1 + corpus | category		intercept, $N = 15$, variance: 1.7066 (intercept), 1.2595 (F-LOB), 1.1406 (Frown), 1.2433 (LOB)			
SUMMARY STATISTICS	*N*	13,192			
	correctly predicted	83.4% (baseline: 55.8%)			
	Somers's D_{xy}	0.84			
	C	0.92			

Table 4. Minimal adequate logistic mixed-effects regression model for variation between *that* and *which* in subject- and nonsubject-RRCs (model 2). The predicted value is *that*.

These findings are helpfully contextualized by considering which of the values for each variable acts as a formal stylistic variant and which is located on the informal, more colloquial end of the spectrum. As was established in the introduction, relativizer *that* is a more colloquial choice relative to *which*. *That* correlates with higher incidences of active-voice verbs, which are more frequent in informal, colloquial contexts of use than in formal language. Thus, both rules prescribe stylistic choices that represent the more informal, colloquial of two options. In the case of preposition stranding, by contrast, prescriptivism calls for the more bookish, formal option: while stranding is common in informal usage contexts, the precept goes against it. This pattern of correlations between observance of the *that*-rule and the other precepts illustrates that the stylistic distribution of relativizer choices is very much involved in the ongoing change.

Furthermore, we note that the interaction term **stranding × time**, like **stranding**, is significant (at $p \cong 0.000$) and makes a prediction supporting the overall hypothesis with regard to change over time: the coefficient for the value '1990s' is negative. This prediction is to be read as follows: for each increase in the percentage of stranded prepositions in a sample, RRCs in texts from 1990 are only 66% as likely to show relativizer *that* as those in texts written in the 1960s. In other words, over the thirty-year time span

covered by the corpora, writers adjust their practice toward less divergent handling of the stranding rule and the *that*-rule.

REGISTER EFFECTS IN THE UPTAKE OF THE *that*-RULE. The degree of frequency increase for *that* and decrease for *which* in the corpora differs markedly among register categories. The fifteen text types sampled in the Brown corpus design fall into four meta-genres (see appendix): press, general prose, learned, and fiction. Our data set is coded both for the fifteen-level distinction **category** and for the four-level factor **genre**, in which **category** is nested. Including **category** in either of our models as a fixed effect would have produced overparametrization problems, and so we used the simpler **genre** as a main effect and in interactions. Nonetheless, we included **category** in the random-effect structure for our models (as laid out in §4.2) in order to control for any bias introduced by text category over and above the variation captured by **genre**.

We base our consideration of register variation on three analyses: a simple distributional analysis of *that* vs. *which* across all subject-RRCs in the data set (Fig. 2b), and the model terms for **variety** × **genre** in model 2 (Table 4). Turning first to the distribution shown in Fig. 2b, we find the highest levels of *that*-usage in fiction texts (except in Frown, where *that* is even more popular in press texts); see Figure 4. Even at the earlier time point, writers chose *that* in more than 50% of all cases, illustrating the greater incidence of colloquial language in fiction relative to the more strictly redacted genres (with regard to style) of press and academic writing. Increase of *that* over time is, accordingly, steeper in those genres that are subject to the treatment of heavier editorial hands than fiction.

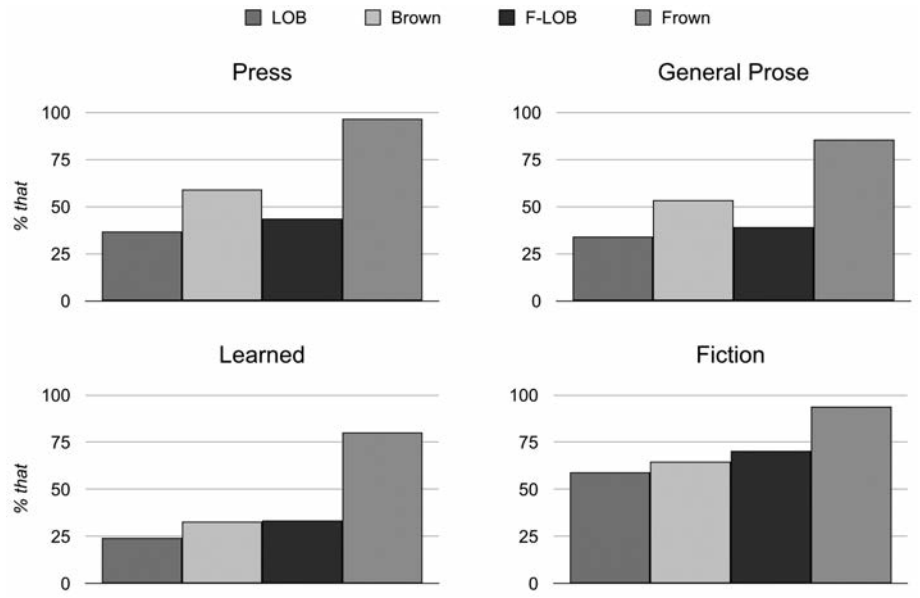


FIGURE 4. Percentage of *that*-choices out of (N_{that} + N_{which}) by corpus and genre. Total N = 13,192: 3,376 (LOB) + 3,090 (Brown) + 3,322 (F-LOB) + 3,404 (Frown).

The 'learned' genre shows the lowest rates of *that*-usage by the end of the coverage period. This observation reflects the orientation of academic prose toward highly formal style, which prefers *which* over *that*. However, in contradiction to Hundt and Mair's (1999) description of press language as 'agile' and learned writing as 'uptight',

the academic texts show the most dramatic rates of increase over time, especially in the American subset: Brown showed only 32.75% *that*, but Frown shows a jump to 80.34%. The case of RRCs, then, sees the academic genre emerging as an interesting battleground between the genre's orientations toward formality on the one hand and what copyeditors consider to be good style on the other.

6. Conclusion. Our analysis has confirmed the findings from earlier studies, such as Leech et al. 2009, which showed that relativization strategies in written, edited, and published StE are changing: relativizer *that* is used significantly more frequently in corpora from the 1990s than in material from the 1960s; this change is happening at the expense of the other overt option *which* and is leaving *zero* largely unaffected.

Against this backdrop, the fact that in our data *that* is on the rise has limited news value. The more interesting questions that this article has sought to address are: Why is *that* on the rise? Is prescriptivist discourse implicated in this development? Our point of departure was that a mere discussion of usage frequencies cannot satisfactorily, we believe, link the increasing popularity of *that* to recommendations in style guides. Instead we suggested that an analysis of the conditioning of *that*, along with probabilistic links to other features targeted by prescriptivists, can yield more conclusive evidence. In this spirit we used multivariate modeling and included a novel set of independent variables, which we labeled 'prescriptivism-related factors'. In the analysis, the choice of *that* over rival options was modeled as dependent on, among other things, each writer's degree of uptake for various areas of variation in StE grammar that are subject to the recommendations of prescriptive usage guides. Among these rules, the ones relating to voice ('use verbs in the active voice rather than passive') and preposition placement ('avoid stranding') emerged as making statistically significant predictions as to the choice of relativizer—and in converse directions: texts that more diligently adhere to the no-stranding rule are overall less likely to follow the *that*-rule in restrictive relatives; that is to say, it is predominantly a text's overall formality level that influences whether writers prefer the formal option for each variable (*which*/no preposition stranding) or the informal option (*that*/stranding). However, the degree of this alignment in writers' and editors' orientation toward both variables according to formality lessens over time, indicating that by the 1990s, the prescription in favor of *that* has turned it into a desirable choice even in more formal writing, where it now increasingly cooccurs with lower frequencies of preposition stranding. Meanwhile, those texts that are in closer alignment with the no-passives rule are significantly more likely to obey the *that*-rule. This alignment process is clearly helped along by the fact that the prescriptive rule, in both cases, advocates the more informal of two available choices.

Thus we have provided circumstantial evidence on the nature of the influence of prescriptive grammar on language use: we are convinced that it does exert a potentially strong influence on the linguistic choices made by actual language users. But rather than spreading evenly as a monolithic force throughout the morphosyntactic system of StE, prescriptivism operates at multiple sites at once, and the effects of the various prescriptive rules that we have considered follow trajectories that are quite distinct from one another.

On the methodological plane, we have demonstrated that the inclusion of a portfolio of prescriptivism-influenced alternations, operationalized quantitatively as independent variables, is one way of approximating a measure of the influence of prescriptive language ideology on usage. The utility of such a method is, of course, not restricted to this specific context, but applies to any study that is interested in the coherence of multiple

linguistic variables. In a recent article, Guy (2013) finds that it is not always empirically straightforward to demonstrate correlations in the behavior of variables, even if they are generally thought to vary along the same social dimension. One key problem is that any given variable is rarely stratified along only one dimension but typically carries multivalent indexical potential. For instance, once Guy (2013:69–70) includes gender and age in addition to social prestige in his analysis of four variables in Brazilian Portuguese, new systematicities in their distribution become discernible.

The analysis we have presented here is sensitive to variable coherence for two reasons. First, we interpret variable behavior not with regard to global characteristics of an individual such as age, ethnicity, or socioeconomic status, but instead draw on aspects of the relatively immediate context (the 2,000-word sample in which the instance of the variable occurs and its text-linguistic properties) for statistical analysis. This allows us to model the actual cooccurrence of variants in contiguous stretches of discourse rather than introduce a relatively abstract mediating category located at the level of the speaker (note, for instance, that in Guy's study, each speaker was interviewed on several distinct occasions, but the individual interviews are not utilized as factors in the analysis). Second, the multivariate statistical modeling we employ allows us to control for a large number of predictors. We are thus able to interpret coherence between two variables while keeping all other predictors equal. In such an analysis, the fact that a variable is stratified along multiple dimensions does not present a problem as long as those dimensions enter the model as predictors.

Meanwhile, we remain doubtful as to the role of grammar checkers in the shift toward *that*. It is true that automatic grammar checkers do enforce the *that*-rule, and popular word processors such as Microsoft Word started implementing grammar checkers in the 1990s. While it so happens that Frown is the corpus in our study showing the farthest advancement of the change, in the form of the highest rates of *that*-usage, and it is also the one from the most recent sampling period—1992, as opposed to 1991 for F-LOB (Hinrichs et al. 2010:200)—word-processing software with integrated grammar checkers was not commonly available until 1997.[17] We conclude therefore that prescriptive grammarians were successful at planting the *that*-rule firmly within British and American English editorial practice by the early 1990s. The 'recommendations' of mass-market grammar checkers, integrated into commonly available word processors such as Microsoft Word, are currently driving the last nail into the coffin of *which* as restrictive relativizer in written StE.

The observed change in favor of *that* shows some of the formal qualities of AMERICANIZATION (Mair 2006:193)—in that it is farther advanced in AmE than in BrE—and COLLOQUIALIZATION (Mair 2006:183)—because *that* is doubtless the more colloquial and vernacular option vis-à-vis bookish *which*. In this way, the change in progress examined here resembles other well-documented shifts that have been explained in the broader context of colloquialization and Americanization, for example, the shift from canonical modal verbs (e.g. *must*) to so-called 'emerging modals' (Krug 2000, e.g. *need to*, *want to*/*wanna*). But the staggering rate at which *that* has been taking over the RRC, along with the observed increase over time in writers' compliance with other prescriptive rules, suggests that the ideas of prescriptive grammarians, handed down through usage

[17] Grammar checkers such as Houghton-Mifflin's CorrecText, which was later incorporated into Microsoft Word, started to be sold as stand-alone software as early as in the 1980s (Dobrin 1990). But until they became part of word-processing software, these programs would have played no role in the writing practice of authors, and a minor role compared to the practice of manual editing in the work of publishing houses.

guides and the educational systems of the English-speaking world, had the effect of massively reinforcing and accelerating an existing trend in the direction of colloquialization. With prescriptivism-based editorial practice and the overall trend of colloquialization blowing, in a rare show of unity, very much into the same horn, one can indeed speak of the recent change in the standard written English relative clause reported here as a case of AmE-led colloquialization with institutional backing. To use a term proposed by Christian Mair (p.c.): the increase of *that* in British and American StE is a case of 'colloquialization from above', remarkable because 'changes from above', which in the customary understanding of Labovian sociolinguistics originate 'above' the level of consciousness and frequently have the benefit of overt, positive metalinguistic discourse to propel them along, are typically directed toward formal prestige variants. In this case, however, the infrastructure of prestige, including the educational system and editorial practice, are helping along a change toward an informal variant.

A third tendency that has been invoked to explain some of the ongoing morphosyntactic changes documented in the Brown corpora, 'densification' (Leech et al. 2009) or 'economization' (Hinrichs & Szmrecsanyi 2007), is not at the root of this change. While one of the relativizer options we studied (*zero*) is indeed more economical than the other two options (*which* and *that*), our multivariate approach showed that the process of relativizer deletion (or: choice of *zero*), unlike selection among overt forms, is unaffected by the prescriptivism-related factors, and rather than undergoing any remarkable change over time, it is conditioned in very similar ways in the 1960s and in the 1990s data.

In conclusion, this study suggests that the rise of *that* in written StE is indeed linked to prescriptivist *which*-hunting and accelerated by the fact that the precept calls for a form whose discourse function is traditionally that of the more colloquial option, relative to alternatives. In order to gauge more fully the effect of this prescriptivism-driven change in StE on the broader discourse norms of English speakers, it will be necessary in future work to study not only the conditioning of relativizer choices in spoken English—some studies in this vein are already available—but also the ways in which English speakers perceive and judge choices of different relativizers. In any case, we conclude that the editorial practice known as *which*-hunting has narrowed, if not eliminated, the gap between written StE and spoken English, where *that* is and always has been the default option.

Appendix: Corpus design of the Brown family

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

ALGEO, JOHN. 1991. Sweet are the usages of diversity. *Word* 42.1.1–17. DOI: 10.1080/00437956.1991.11435829.

ANDERWALD, LIESELOTTE. 2011. Norm vs variation in British English irregular verbs: The case of past tense *sang* vs *sung*. *English Language and Linguistics* 15.1.85–112. DOI: 10.1017/S1360674310000298.

AUER, ANITA. 2006. Precept and practice: The influence of prescriptivism on the English subjunctive. *Syntax, style and grammatical norms: English from 1500–2000* (Studies in language and communication 39), ed. by Christiane Dalton-Puffer, Dieter Kastovsky, Nikolaus Ritt, and Herbert Schendl, 35–53. Berne: Peter Lang.

AUER, ANITA, and VICTORINA GONZÁLEZ-DÍAZ. 2005. Eighteenth-century prescriptivism in English: A re-evaluation of its effects on actual language usage. *Multilingua* 24.4. 317–41. DOI: 10.1515/mult.2005.24.4.317.

BAAYEN, R. HARALD. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

BALL, CATHERINE N. 1996. A diachronic study of relative markers in spoken and written English. *Language Variation and Change* 8.2.227–58. DOI: 10.1017/S0954394500001150.

BATES, DOUGLAS M.; MARTIN MAECHLER; and BEN BOLKER. 2011. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. Online: http://cran.r-project.org/package=lme4.

BIBER, DOUGLAS. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

BIBER, DOUGLAS. 2003. Compressed noun-phrase structures in newspaper discourse: The competing demands of popularization vs. economy. *New media language*, ed. by Jean Aitchison and Diana M. Lewis, 169–81. London: Routledge.

BIBER, DOUGLAS, and EDWARD FINEGAN. 2001. Diachronic relations among speech-based and written registers in English. *Variation in English: Multidimensional studies*, ed. by Susan Conrad and Douglas Biber, 66–83. London: Longman.

BIBER, DOUGLAS; STIG JOHANSSON; GEOFFREY LEECH; SUSAN CONRAD; and EDWARD FINEGAN. 1999. *The Longman grammar of spoken and written English*. London: Longman.

BLOOMFIELD, LEONARD. 1927. Literate and illiterate speech. *American Speech* 2.10.432–39. DOI: 10.2307/451863.

BOHMANN, AXEL, and PATRICK SCHULTZ. 2011. Sacred that and wicked which: Prescriptivism and change in the use of English relativizers. *Proceedings of the Nineteenth Annual Symposium About Language and Society–Austin*. Online: http://studentorgs.utexas.edu/salsa/proceedings/2011/09TLF54-BohmannSchultz.pdf.

BUSSE, ULRICH, and ANNE SCHRÖDER. 2006. From prescriptivism to descriptivism? 140 years of English usage guides: Some old and new controversies. *Anglistentag 2005 Bamberg: Proceedings*, ed. by Christoph Houswitschka, Gabriele Knappe, and Anja Müller, 457–73. Trier: WVT.

BUSSE, ULRICH, and ANNE SCHRÖDER. 2010. Problem areas of English grammar between usage, norm, and variation. *Grammar between norm and variation* (VarioLingua 40), ed. by Alexandra M. Lenz and Albrecht Plewnia, 87–102. Frankfurt: Peter Lang.

CAMERON, DEBORAH. 1995. *Verbal hygiene: The politics of language*. London: Routledge.

CHAPMAN, DON. 2010. Bad ideas in the history of English usage. *Variation and change in English grammar and lexicon: Contemporary approaches* (Studies in the history of the English language 5), ed. by Robert Cloutier, Anne Marie Hamilton-Brehm, and William Kretzschmar, 141–60. Berlin: Mouton de Gruyter.

CHESHIRE, JENNY. 1987. Syntactic variation, the linguistic variable, and sociolinguistic theory. *Linguistics* 25.2.257–82. DOI: 10.1515/ling.1987.25.2.257.

CURZAN, ANNE. 2014. *Fixing English: Prescriptivism and language change*. Cambridge: Cambridge University Press.

D'ARCY, ALEXANDRA, and SALI TAGLIAMONTE. 2010. Prestige, accommodation, and the legacy of relative *who*. *Language in Society* 39.3.383–410. DOI: 10.1017/S0047404510000205.

DENISON, DAVID, and MARIANNE HUNDT. 2013. Defining relatives. *Journal of English Linguistics* 41.2.135–67. DOI: 10.1177/0075424213483572.

Dobrin, David N. 1990. A new grammar checker. *Computers and the Humanities* 24.1–2.67–80. DOI: 10.1007/BF00115029.

Facchinetti, Roberta. 2000. The modal verb *shall* between grammar and usage in the nineteenth century. *The history of English in a social context: A contribution to historical sociolinguistics*, ed. by Dieter Kastovsky, 115–33. Berlin: Mouton de Gruyter.

Finegan, Edward, and Douglas Biber. 1994. Register and social dialect variation: An integrated approach. *Sociolinguistic perspectives on register*, ed. by Edward Finegan and Douglas Biber, 315–47. Oxford: Oxford University Press.

Fischer, Olga. 1992. Syntax. *The Cambridge history of the English language, vol. 2: 1066–1476*, ed. by Richard Hogg and Norman Blake, 207–408. Cambridge: Cambridge University Press.

Fowler, Henry W., and David Crystal. 2009. *A dictionary of modern English usage*. Oxford: Oxford University Press.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gorman, Kyle, and Daniel E. Johnson. 2013. Quantitative analysis. *The Oxford handbook of sociolinguistics*, ed. by Robert Bayley, Richard Cameron, and Ceil Lucas, 214–40. Oxford: Oxford University Press.

Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34.365–99. DOI: 10.1007/s10936-005-6139-3.

Gries, Stefan Th., and Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14.3.293–320. DOI: 10.1017/S1360674310000092.

Guy, Gregory R. 2013. The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics* 52.63–71. DOI: 10.1016/j.pragma.2012.12.019.

Guy, Gregory R., and Robert Bayley. 1995. On the choice of relative pronouns in English. *American Speech* 70.2.148–62. DOI: 10.2307/455813.

Haussamen, Brock. 1997. *Revising the rules: Traditional grammar and modern linguistics*. Dubuque, IA: Kendall/Hunt.

Hinrichs, Lars; Nicholas Smith; and Birgit Waibel. 2010. Manual of information for the part-of-speech-tagged, post-edited 'Brown' corpora. *ICAME Journal* 34.189–231. Online: http://clu.uni.no/icame/ij34/F-LOB_Frown_manual.pdf.

Hinrichs, Lars, and Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11.3.437–74. DOI: 10.1017/S1360674307002341.

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

Hundt, Marianne. 2009. Colonial lag, colonial innovation or simply language change? *One language, two grammars? Differences between British and American English*, ed. by Günther Rohdenburg and Julia Schlüter, 13–37. Cambridge: Cambridge University Press.

Hundt, Marianne; David Denison; and Gerold Schneider. 2012. Retrieving relatives from historical data. *Literary and Linguistic Computing* 27.1.3–16. DOI: 10.1093/llc/fqr049.

Hundt, Marianne, and Christian Mair. 1999. 'Agile' and 'uptight' genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4.221–42. DOI: 10.1075/ijcl.4.2.02hun.

Jaeger, T. Florian. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Stanford, CA: Stanford University dissertation. Online: http://www.bcs.rochester.edu/people/fjaeger/thanks.pdf.

Johansson, Stig, and Knut Hofland. 1989. *Frequency analysis of English vocabulary and grammar: Based on the LOB corpus*. Oxford: Clarendon.

Johnson, Daniel E. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3.1.359–83. DOI: 10.1111/j.1749-818X.2008.00108.x.

Keenan, Edward L., and Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8.1.63–99.

KOCH, PETER, and WULF OESTERREICHER. 1985. Sprache der Nähe—Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 39.15–43. DOI: 10.1515/9783110244922.15.

KRUG, MANFRED G. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: De Gruyter.

KUČERA, HENRY, and W. NELSON FRANCIS. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

LABOV, WILLIAM. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45.4.715–62. DOI: 10.2307/412333.

LABOV, WILLIAM. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

LAFFERTY, JOHN; ANDREW MCCALLUM; and FERNANDO PEREIRA. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML-01*, 282–89.

LEECH, GEOFFREY; MARIANNE HUNDT; CHRISTIAN MAIR; and NICHOLAS SMITH. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

LEECH, GEOFFREY, and NICHOLAS SMITH. 2006. Recent grammatical change in written English 1961–1992: Some preliminary findings of a comparison of American with British English. *The changing face of corpus linguistics*, ed. by Antoinette Renouf and Andrew Kehoe, 185–204. Amsterdam: Rodopi.

LEECH, GEOFFREY, and NICHOLAS SMITH. 2009. Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931–1991. *Corpus linguistics: Refinements and reassessments*, ed. by Antoinette Renouf and Andrew Kehoe, 173–200. Amsterdam: Rodopi.

LEHMANN, HANS-MARTIN. 2002. Zero subject relative constructions in American and British English. *New frontiers of corpus research: Papers from the twenty-first international conference on English language research on computerized corpora, Sydney 2000* (Language and computers: Studies in practical linguistics 36), ed. by Pam Peters, Peter Collins, and Adam Smith, 163–78. Amsterdam: Rodopi.

LEVEY, STEPHEN. 2006. Visiting London relatives. *English World-Wide* 27.1.45–70. DOI: 10.1075/eww.27.1.04lev.

MAIR, CHRISTIAN. 2006. *Twentieth-century English: History, variation, and standardization*. Cambridge: Cambridge University Press.

MAIR, CHRISTIAN, and GEOFFREY LEECH. 2006. Current changes in English syntax. *The handbook of English linguistics*, ed. by Bas Aarts and April McMahon, 318–42. Malden, MA: Blackwell.

MARCUS, MITCHELL P.; MARY ANN MARCINKIEWICZ; and BEATRICE SANTORINI. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19.2.313–30. Online: http://www.aclweb.org/anthology/J93-2004?CFID=547067514&CFTOKEN=12393887.

NEVALAINEN, TERTTU. 2012. Reconstructing syntactic continuity and change in Early Modern English regional dialects: The case of *who*. *Analysing older English*, ed. by David Denison, Ricardo Bermúdez-Otero, Chris McCully, and Emma Moore, 159–84. Cambridge: Cambridge University Press.

OLOFSSON, ARNE. 1981. *Relative junctions in written American English*. Gothenburg: University of Gothenburg.

PETERS, PAM, and WENDY YOUNG. 1997. English grammar and the lexicography of usage. *Journal of English Linguistics* 25.4.315–31. DOI: 10.1177/007542429702500406.

PICKERING, MARTIN J., and HOLLY P. BRANIGAN. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language* 39.633–51. DOI: 10.1006/jmla.1998.2592.

POPLACK, SHANA, and NATHALIE DION. 2009. Prescription vs. praxis: The evolution of future temporal reference in French. *Language* 85.3.557–87. DOI: 10.1353/lan.0.0149.

PULLUM, GEOFFREY K. 2009. 50 years of stupid grammar advice. *The Chronicle of Higher Education*, April 17, The Chronicle Review 55.32.B15. Online: http://chronicle.com/article/50-Years-of-Stupid-Grammar/25497.

Pullum, Geoffrey K. 2014. Fear and loathing of the English passive. *Language and Communication* 37.60–74. DOI: 10.1016/j.langcom.2013.08.009.
Quirk, Randolph. 1957. Relative clauses in educated spoken English. *English Studies* 38.1–6.97–109. DOI: 10.1080/00138385708596993.
Quirk, Randolph; Sidney Greenbaum; Geoffrey Leech; and Jan Svartvik. 1972. *A grammar of contemporary English*. London: Longman.
R Development Core Team. 2013. R: A language and environment for statistical computing. Version 3.0.1. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org/.
Rickford, John R. 2011. Relativizer omission in Anglophone Caribbean Creoles, Appalachian, and African American Vernacular English [AAVE], and its theoretical implications. *Language from a cognitive perspective: Grammar, usage, and processing: Studies in honor of Thomas Wasow* (CSLI lecture notes 201), ed. by Emily Bender and Jennifer Arnold, 139–60. Stanford, CA: CSLI Publications.
Riley, Kathryn, and Frank Parker. 1998. *English grammar: Prescriptive, descriptive, generative, performance*. Boston: Allyn & Bacon.
Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7.149–82. DOI: 10.1515/cogl.1996.7.2.149.
Romaine, Suzanne. 1982. *Socio-historical linguistics: Its status and methodology*. Cambridge: Cambridge University Press.
Sand, Andrea, and Rainer Siemund. 1992. LOB—30 years on … . *ICAME Journal* 16. 119–22.
Sankoff, David. 1989. Sociolinguistics and syntactic variation. *Linguistics: The Cambridge survey*. *Vol. 4: Language: The socio-cultural context*, ed. by Frederick J. Newmeyer, 140–61. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511620577.009.
Strunk, William, and E. B. White. 1999. *The elements of style*. 4th edn. London: Longman.
Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: Mouton de Gruyter.
Szmrecsanyi, Benedikt, and Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English. *The dynamics of linguistic variation: Corpus evidence on English past and present*, ed. by Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta, and Minna Korhonen, 291–309. Amsterdam: John Benjamins.
Tagliamonte, Sali. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley-Blackwell.
Tagliamonte, Sali, and R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24.2.135–78. DOI: 10.1017/S0954394512000129.
Tagliamonte, Sali; Jennifer Smith; and Helen Lawrence. 2005. No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change* 17.1.75–112. DOI: 10.1017/S0954394505050040.
Tottie, Gunnel, and Dawn Harvie. 2000. It's all relative: Relativization strategies in early African American Vernacular English. *The English history of African American English*, ed. by Shana Poplack, 198–230. Oxford: Blackwell.
Tweedie, Fiona J., and R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32.323–52. DOI: 10.1023/A:1001749303137.
Weisberg, Sanford, and John Fox. 2011. *An R companion to applied regression*. London: Sage.
Wolfram, Walt. 1991. The linguistic variable: Fact and fantasy. *American Speech* 66.1.22. DOI: 10.2307/455432.
Wolk, Christoph; Joan Bresnan; Anette Rosenbach; and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30.3.382–419. DOI: 10.1075/dia.30.3.04wol.

WOOLARD, KATHRYN A., and BAMBI B. SCHIEFFELIN. 1994. Language ideology. *Annual Review of Anthropology* 23.55–82. DOI: 10.1146/annurev.an.23.100194.000415.

[larshinrichs@utexas.edu]
[benszm@kuleuven.be]
[axel.bohmann@mail.utexas.edu]

[Received 27 March 2014;
accepted 13 June 2014]