Restricting the restrictive relativizer

Constraints on subject and non-subject English relative clauses

Jason Grafmiller, Benedikt Szmrecsanyi, and Lars Hinrichs

Abstract

We investigate internal and stylistic factors affecting binary and ternary relativizer choice in subject (*that vs. which)* and non-subject *(that* vs. *which* vs. *zero*) relative clauses. We employ a novel methodological approach to predicting relativizers: Bayesian regression modeling with the dimensional reduction of model inputs via factor analysis. Our factor analysis is motivated by the high degree of redundancy and collinearity in natural language data, while Bayesian regression models are robust to effects of data sparseness and (near) separation. We find that in both types of relative clauses, the more marked variant (*which*) is preferred in complex contexts, while the unmarked variant (*that*, or *zero* in NSRCs) is favored in contexts where the relative clause is short and more fully integrated with the NP it modifies. We also find that use of *which* is somewhat more sensitive to stylistic considerations in subject than in non-subject relative clauses, and that *which* correlates most strongly with features associated with lexical density, e.g. 'nouniness', rather than those often associated with formality, e.g. passivization and sentence length.

Keywords: English relativizers, complexity, stylistic variation, Bayesian modelling, multinomial regression, factor analysis

1 Introduction

When introducing a relative clause, users of standard written English (StE) may choose from a number of available syntactic markers. These include a variety of so-called *wh-* forms, e.g. *which*, *who*, *whom*, and *whose*, as well as *that*, and even no pronoun at all (*'zero'* or ∅).

(1) *The landed gentry indeed led the way in such investment in a way* ***which*** *did not happen in any other European country.* <F-LOB:G34>

(2) *Those party to the consensus hold, among other things, that if the best* ***that*** *can be said for a belief is that it is their belief,…* <F-LOB:J52>

(3) *… that what we were asserting to be bad was precisely the suffering* ***∅*** *we thought had occurred back there…* <Brown:J52>

Studies of the English relativizer system have investigated variation in standard spoken and written varieties (e.g. Guy and Bayley 1995; Lehmann 2001; Hinrichs, Szmrecsanyi and Bohmann 2015) and in vernacular speech (e.g. Tottie and Harvie 2000; Tagliamonte, Smith and Lawrence 2005; Levey 2006), from both synchronic and diachronic perspectives (e.g. Romaine 1980; Ball 1996; Hundt, Denison and Schneider 2012; Nevalainen 2012). In addition, relativizer variation has proven to be a useful phenomenon for exploring the relation between language processing and syntactic structure (Rohdenburg 1996; Temperley 2003; Fox and Thompson 2007; Levy and Jaeger 2007; Gennari and MacDonald 2009; Wells et al. 2009; Jaeger 2011; Wiechmann 2015; among others).

A wide range of factors influence the choice from among these different options, several of which function (nearly) categorically in present-day StE. Of these factors, perhaps the most elementary is the distinction between restrictive or non-restrictive relative clauses. Only *wh-*forms may introduce a non-restrictive relative clause (Quirk 1957), and given this constraint, most studies of relativizer variation limit their investigation to restrictive relative clauses. A second, relatively recent constraint targets the animacy of the antecedent: animate antecedents overwhelmingly prefer *who(m/se)* while the use of *that* is available with both animate and inanimate antecedents (Ball 1996; Nevalainen 2012). The use of *that* with animate antecedents is however quite rare in written English (Biber et al. 1999:613), but this pattern does not hold in spoken (vernacular) English, where the use of *that* with animate antecedents is common whereas use of *which* is quite rare (D’Arcy and Tagliamonte 2010, 2015). Both the restrictiveness and animacy constraints have long been targets of prescriptive advice (e.g. Fowler and Crystal 2010), and the stylistic variability of the different options in written StE is well-known (see Biber et al. 1999; Leech et al. 2009). For these reasons, we isolate our investigation in the present study to restrictive relative clauses (RCs) with inanimate antecedents to focus on the choice between *that, which,* and *zero* in RCs in the Brown corpora of British and American standard written English.

A third major factor governing relativizer choice takes center stage in this paper, namely the syntactic status/function of the relativizer inside the RC, i.e. subject (4) or non-subject (5).

(4) *Two issues concerned payroll practices* ***that*** *created an underpayment of man-hour contributions*. <Frown:J72>

(5) *’t is plain that the murderer wrapped his ill-gotten gains in the first thing* ***that*** *he could find and placed it in our thatch*. <LOB:L15>

At issue is the fact that one of the options available in non-subject-extracted RCs, *zero*, is not available with subject-extracted RCs in StE. Aside from simply noting the categorical lack of *zero* in subject RCs, few studies of relativizer variation have addressed the importance of relativizer function head on (cf. Fox and Thompson 2007; Wiechmann 2015). There is a long tradition of experimental research demonstrating that non-subject RCs tend to be more difficult to process than subject RCs (see Gennari and MacDonald 2009 for review), but we are not aware of any such studies that include the form of the (overt) relativizer itself as an independent variable. Recent research has nonetheless demonstrated that processing and comprehension demands undoubtedly play a role in shaping relativizer choice (e.g. Rohdenburg 1996; Jaeger 2011), but again, this work has tended to focus solely on relativizer omission while ignoring choices among overt variants (though cf. Rohdenburg 2014). Studies that examine the full envelope of relativizer variation in StE in terms of stylistic demands as well as internal cognitive pressures are few and far between.

The aim of this paper is to bring together these various strands of research to develop a more comprehensive understanding of the relativizer system in present-day standard written English. The basic question we are investigating is *to what extent do the same factors that affect relativizer choice in subject-extracted relative clauses (SRCs) also affect relativizer choice in* *non-subject-extracted relative clauses (NSRCs)?* We expand upon recent studies of restrictive relativizers by considering the role of internal and stylistic pressures on relativizer choice among all three response categories *that*, *which*, and *zero*. The simultaneous investigation of SRCs and NSRCs presents a number of challenges, not least of which is the fact that certain options, i.e. *zero*, are available in one context and not the other. In addition, many factors known to influence relativizer choice are highly correlated with one another, which can lead to potentially spurious inferences about their individual effects. We confront these challenges through the use of statistical methods that have heretofore been under-utilized in variationist research, namely Bayesian multinomial modeling and the dimension reduction of multivariate predictors (cf. Levshina 2016). We find that on the whole, variation in both SRC and NSRC contexts is largely shaped by the same forces, albeit in subtly different ways. In particular, we find that the tendency in both SRCs and NSRCs is toward the use of the more explicit or marked variant (*which*) in relatively complex contexts, and toward use of the unmarked variant (*that*, or *zero* in NSRCs) in contexts where the RC is more fully integrated with the NP it modifies, and thus arguably more likely given the preceding material. At the same time, we find that use of *which* is more sensitive to stylistic considerations in SRCs than in NSRCs, and that its use does not necessarily correlate with features most often associated with high degrees of formality. We argue that these findings have import not only for theories of the English relativizer system, but for our understanding of the nature of syntactic variation in general.

2 Background

In this section we briefly review the current state-of-the-art in our understanding of English relativizer variation. Since the present study explores variation in present-day written StE, we focus our discussion on relevant factors within that domain, while acknowledging the rich tradition of research on the English relativizer system from historical and/or sociolinguistic perspectives (Ball 1996; Tottie and Harvie 2000; Rickford 2011; among many others).

2.1 Relativizer variation from an external perspective: time, region, and style

Setting aside clauses with animate antecedents, present-day StE relatives are characterized by a clear stylistic difference among *that*, *which*, and *zero* as restrictive relativizers. According to

Biber et al. (1999:610), *that* is the most frequently used relativizer in spoken, conversational language, followed in frequency by *zero* and finally *which*. As for written StE, academic prose differs most strongly from conversational speech in this respect: *which* is the most frequently used relativizer in this register, with *that* in second place and *zero* in third (Biber et al. 1999:611). The distributions in other written registers vary, with fiction texts resembling conversational speech slightly more than news writing does. It is often inferred from these register preferences the informal stylistic value of *that* (see also D'Arcy and Tagliamonte 2010) and the formal value of *which*. *That* is preferred in more informal, "involved" style, while *which* is more likely to be found in formal or "informational" register/texts.[1] Drawing on data from 20th century StE, Hinrichs et al. (2015) demonstrate that use of *which* correlates with various features associated with more elevated style in both SRCs and NSRCs, including an overall greater likelihood in academic writing (see also Leech et al. 2009:226–233). As Hinrichs et al. note, *which* and *that* have held these stylistic connotations ever since *which* was first used as a relativizer in the Late Middle English period (Fischer 1992:296). By contrast, use of *zero* shows much less variability across registers than the two overt forms, thus its primary stylistic value is much less clear.

In terms of regional variation, American English shows greater use of *that* than British English in both SRCs and NSRCs. This bias toward *that* in written American English is a quite recent development, having risen dramatically only in the second half of the twentieth century (Hundt, Denison and Schneider 2012). In NSRCs this change is coming almost entirely at the expense of *which* (Leech et al. 2009), and though the trend was first noted only in 20th century American English texts, evidence from 21st century data suggests that written British English also appears to be changing in a similar direction (Hundt and Leech 2012). In some ways this development among relativizers reflects a general trend toward colloquialization in written American English,

[1] Note that this sense of 'information' as used in register variation studies should not be confused with the notion of (Shannon) information that is central to information theory and information-theoretic approaches to language processing.

though the reasons for its more recent development in written British English remain unclear. Still, the degree to which *that* is overtaking *which*, even in formal texts, is quite striking, and in need of further explanation. One hypothesis attributes these changes to the prescriptive prohibition against restrictive *which* in the United States (Leech et al. 2009; Hundt and Leech 2012). In a recent study, Hinrichs et al. (2015) demonstrate clear correlations between increased use of *that* vs. *which* and use of other prescriptive strictures in the Brown corpora of StE, and argue that the relativizer system is undergoing a curious case of 'colloquialization from above' (831). In the present study we build upon Hinrichs et al.'s work in order to take full stock of the role of stylistic pressures on relativizer choice, by including the *zero* option as a third response category.

2.2 Relativizer variation from an internal perspective: processing and cognitive complexity

In addition to stylistic constraints, relativizer variation is highly sensitive to internal linguistic constraints, in particular those that are thought to reflect the influence of deeper cognitive processes on syntactic variation in general. As mentioned above, studies looking at both SRCs and NSRCs have tended to focus on differences in overall processing difficulty between the two classes (e.g. Keenan and Comrie 1977). Recent work explains such processing difficulties in terms of language users' sensitivity to probabilistic regularities in the language (see Gennari and MacDonald 2009 for review). Semantic factors, in particular the animacy of the antecedent, have been shown to mitigate the supposed difficulties of NSRCs. Non-subject RCs with inanimate antecedents are easier to process than those with animate antecedents (e.g. Traxler et al. 2005; Mak, Vonk and Schriefers 2006), and experimental results align with findings from corpora, where inanimate-headed NSRCs are more frequent than animate-headed ones (Roland, Dick and Elman 2007). Similar parallels between patterns in production data (corpora) and performance in on-line comprehension tasks have been demonstrated in other languages as well (e.g. Desmet, De Baecke, Drieghe, Brysbaert and Vonk 2006).

While it is true that different factors may affect production and comprehension in different ways, a certain degree of convergence between the two processes is expected under many usage-based models of grammar, e.g. Bybee (2010). Such approaches maintain that language users implicitly learn statistical tendencies in the distribution of forms they are exposed to, and that the statistical associations among linguistic structures are used to guide production and comprehension, possibly in concert with other cognitive factors. The social dimension of language also plays a role under this view, thus it is predicted that community-specific social forces, e.g. language attitudes or stylistic preferences, also shape biases in production, which are in turn reflected in specific forms' distributions.

We adopt such a model of usage-driven syntactic variation in the present study, following many previous accounts of relativizer choice in English. While the present study considers variation among *zero* and overt relativizers, most studies have tended to focus primarily on relativizer omission in NSRCs, but all accounts generally converge on the same core findings. Omission is preferred in contexts were the constituents involved are relatively short and not complex, and where the probability of a relative clause is high given the material that preceded it. Explanations for these effects come in several flavors.

The 'complexity' approach maintains that the use of more explicit forms is preferred in cognitively complex environments. Overt relativizers signal the presence of an (upcoming) RC, thereby aiding comprehension and making parsing more efficient. The most well-known articulation of this approach is probably Rohdenburg's (1996:151) 'Complexity Principle'.

(6) In the case of more or less explicit grammatical options the more explicit one(s) will tend to be favored in cognitively more complex environments.

The concept of syntactic 'complexity' is itself a thorny issue, but it is most often operationalized in terms of the length of the relevant (noun) phrases (Berlage 2014). As such, the length of the antecedent, and to a lesser extent the RC does indeed influence the choice of relativizer quite reliably in the direction predicted by the Complexity Principle.

When it comes to competition among overt variants, we view 'explicitness' in this context in terms of markedness, and argue that *that* represents the unmarked variant—the "primary [relativization] strategy" in Keenan and Comrie's (1977:67-8) terms—for several reasons. From a phonological/phonetic perspective, *that* involves fewer marked segments than *which*, and these phonemes tend to reduce more phonetically, e.g. /t/ reduces more/has more allophonic variation than /tʃ/. At the same time, *wh-* forms are "semantically more explicit" (Hundt et al. 2012:213), in that they are distinguished by the animacy of their antecedents: *which* unambiguously refers to an inanimate or nonhuman antecedent; *that* is ambiguous. Additionally, the word form *that* serves many different functions (determiner, demonstrative pronoun, complementizer, etc.), hence *that* is overall less informative than *which* as an indicator of an upcoming RC. Finally, we know that *which* is also more frequent in formal discourse domains (Biber et al. 1999), thus *which* is marked by virtue of its stylistic distribution across the entire economy of relativizers in StE.

Beyond complexity approaches, there are accounts which argue that users structure their language so as to avoid temporary ambiguities (e.g. Temperley 2003). Under such approaches, the use of overt forms is preferred in contexts where the potential for garden-pathing or misinterpretation is high. Without an overt marker, SRCs such as (7a) would be ambiguous, while similar NSRCs are not, whether an overt marker is present or not.

(7) a. **The car Ø hit me was red.*

b. The car that hit me was red.

c. The car Ø I hit was red.

d. The car that I hit me was red.

Whether users actually avoid syntactic ambiguities is still an open question however, as recent research finds conflicting evidence for this account (see Jaeger 2011 for discussion).

Alternatively, there are accounts which treat omission in probabilistic terms within an incremental model of language production. A well-known articulation of this approach is the 'Predictability Hypothesis' of Wasow, Jaeger and Orr (2011).

(8) In environments where an NSRC is more predictable, relativizers are less frequent [omission is more common]

The effect of predictability is often interpreted as a strategy for modulating the flow of information (Jaeger 2010, 2011). Such accounts generally assume some version of the Uniform Information Density hypothesis (Levy and Jaeger 2007; Jaeger 2010) which holds that producers shape their language to distribute information uniformly across the linguistic signal. Information of an item in this sense is defined as the log-inverse of its probability, and under this account overt relativizers are predicted in contexts where the information density of the RC would be high if the relativizer were omitted. In other words, *that* is omitted in contexts where it is unnecessary or redundant. Similar to the complexity approaches, the UID model can be seen as a method for minimizing comprehension difficulty, though this model is defined in information-theoretic terms as the surprisal/probability of an RC, given the preceding material.

Finally, there are approaches that view relativizer choice as a function of the degree of entrenchment of the NP (construction) containing the RC (Wiechmann 2015:5), where 'entrenchment' is viewed in cognitive constructionist terms as a routinization process by which highly frequent structures come to be treated (and processed) as a unit (see e.g. Langacker 2008). In this view deep routinization typically leads to formal reduction, hence the greater tendency for omission among highly entrenched structures. The entrenchment view differs from other processing-related accounts, e.g. UID, in that it de-emphasizes the role of online incremental production and focuses on processing effects in the formation of grammar over time through the gradual entrenchment of high-level constructional schemas. While the distinctions between entrenchment and other low-level processing approaches are non-trivial, in practical terms, the predictions made by them overlap considerably (see Wiechmann 2015:215-219).

The common denominator among these different approaches is that the surrounding linguistic material has considerable influence on the choice of relativizer, specifically the choice between *zero* and an overt form. Very few studies however have explored the role of these factors on the choice between overt variants, nor have they examined interaction of these factors with relativizer function. Thus we are left with a number of vexing questions to be addressed in this paper:

- How should 'complexity' affect relativizer choice when an overt marker is required? Assuming that *which* is the more explicit/less reduced variant, is it used more often than *that* in cognitively complex (or less predictable) contexts (see e.g. Rohdenburg 2014)?
- Do we find a division of labor with respect to the three options and the kinds of factors involved? Given that one option for NSRCs is *zero*, do we find that cognitive factors play the biggest role in influencing the choice between something and nothing, while the choice between overt variants is primarily a stylistic one?

3 Data and annotation

3.1 The data

For this study we make use of a modified version of the dataset of StE relativizers analyzed by Hinrichs et al. (2015), which was collected from the four most complete portions of the Brown family of corpora: Brown, Frown, LOB, and F-LOB (Francis and Kucera 1979; Johansson and Hofland 1989; Sand and Siemund 1992). Each corpus contains roughly 1 million words of written text of standard American (Brown, Frown) and British (LOB, F-LOB) English compiled at two distinct time periods over the latter half of the twentieth century: the early 1960s (Brown, LOB) and the early 1990s (Frown, F-LOB). Each corpus consists of 500 2,000-word samples representing data of 15 distinct text categories, e.g. newspaper articles, humor writing, academic writing, and various genres of fiction (see Hinrichs, Smith and Waibel 2010 for further documentation).

3.2 Variable extraction

The present study follows standard methods used in the variationist tradition, focusing on ‘alternate ways of saying “the same” thing’ (Labov 1972:188). We are interested in writers’ choices between *that* and its alternatives in restrictive RCs, thus we limit our attention to only those cases in which both *that* and *which* (or *that*, *which*, and *zero* in NSRCs) are in principle possible. In StE, *who* forms categorically refer to animate antecedents, while *that* and *which* only rarely do so. In our data, animate antecedents account for barely 1% of the observed uses of *that/which*, thus we excluded animate antecedents from our dataset following standard practices (e.g. D’Arcy and Tagliamonte 2015:272-3). Relativizers preceded by a comma were excluded from the dataset, as commas are a reliable diagnostic of non-restrictive RCs in present-day written StE (see Hinrichs et al. 2015:815, fn.6). To preserve interchangeability, we narrowed the variable context by excluding RCs with *wh-* relativizers other than *which*, i.e. *who(m/se)*, oblique

RCs with pied-piping (*that's the house about which I was talking*)[2], and any cases where the relativizer is *when, where*, or *why* and performs an adverbial function. Since relativizer omission is not permitted with subject RCs in written StE but is quite frequent with non-subject RCs, we define two distinct variable contexts: subject-extracted relative clauses (SRCs—*that* vs. *which*) and non-subject-extracted relative clauses (NSRCs—*that* vs. *which* vs. *zero*) (see also Ball 1996).

To extract instances of relative clauses with overt relativizers, Hinrichs et al. (2015) made extensive use of the Part-of-speech tagging available in the four corpora (Hinrichs, Smith and Waibel 2010). A statistical supervised machine learning approach was used to detect *zero*-relatives in the tagged corpora, which offered considerable improvement in both precision and recall (Frazee et al. 2015). For additional details on the extraction methods, see Hinrichs et al. (2015:815–6) and references therein. We note though that the dataset to be analyzed here substantially overlaps with, but is not entirely identical, to the dataset analyzed by Hinrichs et al. Figure 1 shows the distribution of relativizer variants across time and region (N_{total} = 15569: N_{src} = 10285, N_{nsrc} = 5284).

[2] While only *which* is permitted with pied-piping, oblique RCs with stranded prepositions do vary (*that's the house that/which/Ø I was talking about*), hence they were included.

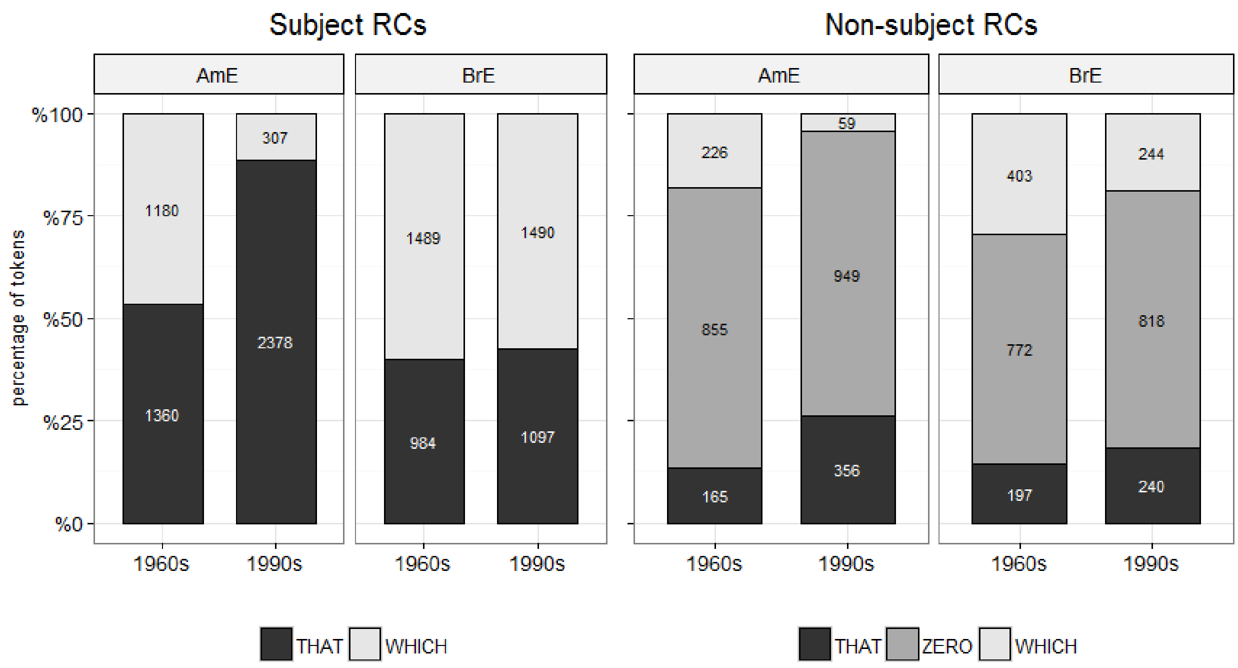


Figure 1: Distribution of relativizers in the Brown corpora, inanimate antecedents only (N = 15569)

3.3 Predictor variables

Each of the 15569 tokens were annotated for a suite of variables thought to influence to the choice of relativizer. These features can be broadly categorized into internal (or linguistic), stylistic, and external predictors. We list each of these below.

3.3.1 Internal predictors

- **Preceding relativizer.** The relativizer used prior to the current one. This includes the levels ‘that’, ‘which’, ‘zero’, and ‘none’ (for when the relativizer in question is the first one encountered in a corpus file). This variable gauges the effect of priming or structural persistence (Szmrecsanyi 2006).
- **Part of Speech of antecedent.** Binary feature of the antecedent head, coded as ‘noun’ vs. ‘other’. This feature gauges the effect of antecedent pronominality independently from definiteness (cf. Tagliamonte, Smith and Lawrence 2005; Tottie and Harvie 2000). This feature also helps distinguish lexically specific antecedents from ‘empty’ ones (e.g. *all* as in

all I said was…) which have been shown to favor *zero* (Fox and Thompson 2007; Wiechmann 2015).

- **Length of antecedent in words.** A measure of the complexity of the noun phrase modified by the relative clause (median: 3 words, interquartile range: 2, 4). We hypothesize that the longer the antecedent, the more likely the use of an overt relativizer becomes (Rohdenburg 1996). For example, the use of an overt variant, e.g. *that*, is heavily favored with an 8 word antecedent as in *the aspects of personal relationships and perceived support that predict recovery in men and women* <F-LOB:J22>.
- **Length of relative clause in words.** A measure of the complexity of the clause introduced by the relativizer (median: 7 words, interquartile range: 3, 10). We hypothesize that the longer the relative clause, the more likely the use of an overt relativizer becomes (Rohdenburg 1996).
- **Adjacency of antecedent head and RC.** A binary indicator (‘Y’/‘N’) of whether the head of the antecedent is adjacent to the left edge of the RC (ignoring the presence/absence of the relativizer). It has been claimed that intervening material favors the use of overt relativizers (e.g. Guy and Bayley 1995), though recent work has failed to confirm this (see Tagliamonte, Smith and Lawrence 2005).
- **Distance in words between the antecedent head and the relative clause.** The number of words between the antecedent head and the relativizer slot. This provides a more detailed, continuous measure of the influence of intervening material on the choice of relativizer. Greater distance between the antecedent and RC is hypothesized to favor overt marking.
- **Number of antecedent.** Grammatical number of the antecedent head, automatically coded as ‘singular’, ‘plural’, or ‘other’ (ambiguous) based on POS-tags. The few studies that have investigated this feature have found that it has only a minor influence on relativizer omission (Tottie and Harvie 2000; Rickford 2011; Hinrichs et al. 2015). We predict nonetheless that whatever influence number may have should be toward greater use of more explicit forms when the antecedent is plural, as plural antecedents are conceptually and formally more complex than singletons (e.g. Clark and Nikitina 2009).
- **Definiteness of the antecedent.** Whether the antecedent was definite (as marked by a definite article, demonstrative, possessive pronoun or genitive *’s*) or indefinite. Indefinite

antecedents have been shown in some studies to favor *zero* (Tagliamonte, Smith and Lawrence 2005; Wasow, Jaeger and Orr 2011), though recent work paints a more equivocal picture for the role of definiteness (Wiechmann 2015).

- **Nested.** A binary indication (‘Y’/‘N’) of whether the RC in question is nested inside another relative clause. To the extent that nested structures are more complex (and thus cognitively more demanding), we expect nested RCs to favor use of the more explicit variant.

3.3.2 Stylistic predictors. The following stylistic features include several “prescriptivism-related” predictors (features 9–12) together with other stylistic features (Hinrichs et al. 2015:818–19).

- **Genre.** Text genre of the sample in which the relativizer occurs. This variable captures four genre meta-categories in the Brown corpora: ‘press’, ‘general prose’, ‘learned’, and ‘fiction’.
- **Type-token ratio (TTR).** The number of unique word types divided by the total number of word tokens in the text sample. Like all measures of lexical density, TTR is not without its limitations (Tweedie and Baayen 1998), however it is a straightforward measure that is commonly used in variationist research (e.g. Hinrichs and Szmrecsanyi 2007; Hinrichs et al. 2015; Szmrecsanyi et al. 2016). Higher TTR values generally predict use of shorter, more compact forms (e.g. Hinrichs and Szmrecsanyi 2007).
- **Personal pronouns.** Frequency of personal pronouns in the corpus text, normalized per 10,000 words. According to Biber (1988), the use of personal pronouns is an indicator of involved style, hence we expect lower personal pronoun use to correlate with greater use of *which*.
- **Nouniness.** Frequency of nouns in the corpus text, normalized per 10,000 words. This provides a measure of the lexical density of a text sample (Leech et al. 2009).
- **Noun to verb ratio.** Ratio between the number of nouns and verbs in the text sample: an alternate measure of lexical density.
- **Mean length of words in text.** 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. Longer average word length is an indicator of greater formality and informational focus, thus we expect use of *which* to increase as average word length increases.

- **Mean length of sentences in text.** The average length in words of a sentence in the text sample. As another indicator of more formal style, this predictor should also be positively correlated with greater use of *which.*
- **Subordinating conjunctions.** Relative frequency of subordinating conjunctions, normalized to a value per 10,000 words in the corpus text under analysis. Increased use of subordinating conjunctions is a reflection of more complex sentence structure and hence yet another measure of formal style.
- **Passives.** Relative frequency of passive constructions, normalized to a value per 10,000 words.
- **Passive to active verb ratio.** Number of passive constructions divided by number of active lexical verbs in the text sample. Together with Passive frequency, this predictor should exert a positive influence on the use of *which,* as frequent passivization is a marker of formal prose. At the same time, correlation between avoidance of passives and use of *that* has been argued to be evidence of the influence of prescriptive traditions (Hinrichs et al. 2015:826).
- **Split infinitives.** Relative frequency of split infinitives (*to boldly go*) in the text sample, normalized to a value per 10,000 words.
- **Stranding.** Proportion of stranded prepositions out of all prepositions in a given corpus text, multiplied by 100 to yield a percentage. Lower frequencies of preposition stranding and split infinitives have been hypothesized to correlate with obedience of other prescriptive maxims, e.g. the exclusive use of *that* with restrictive RCs (Hinrichs et al. 2015).

3.3.3 External predictors. In addition to the internal and stylistic predictors just mentioned, we included two external predictors to control for variation across the two time periods and varieties represented in the Brown corpora. Results pertaining to these predictors can be found in the appendix.

- **Time.** A binary indicator (‘1960s’ vs. ‘1990s’) of the time period when the text was sampled.
- **Variety.** The variety of written StE (‘AmE’ vs. ‘BrE’) from which the text was sampled.

3.4 Multicollinearity and Factor analysis

When the goal of a multivariate analysis of any phenomenon is to make inferences about causal relationships between independent variables and the outcome, the ideal situation is one where the independent variables are uncorrelated with each other. However, this is rarely ever the case in studies of natural language data, as many variables are often highly correlated with one or more other variables—a phenomenon known as multicollinearity. In regression modeling, severe multicollinearity does not affect the predictive power of a model as a whole, however it does present serious problems for interpreting the effects of individual predictors in the model. This is because it is difficult to say how much of the variance in the outcome, e.g. the choice of relativizer, is uniquely captured by the independent predictors, e.g. mean word length and nouniness. Thus, when faced with a model of data involving a high degree of multicollinearity, it becomes very difficult to tease apart the explanatory contribution of individual predictors.

One way to resolve the problem of multiply correlated predictor variables is to transform the original variables via Principle Components Analysis (in the case of continuous predictors) or Multiple Correspondence Analysis (in the case of categorical predictors). Both techniques are designed to transform a set of observations of correlated variables into a set of values of linearly uncorrelated variables, or dimensions, where the number of dimensions is less than or equal to the number of original variables. This transformation is defined in such a way that the first dimension accounts for the largest possible variance in the data, and each successive dimension accounts for less and less variance. The resulting dimensions can be included as predictors in a regression model, and because they are by definition uncorrelated, it is possible to reliably assess their independent effects on the outcome.

Since our features are a mix of categorical and continuous predictors, applying standard methods across the full set(s) of predictors is inappropriate. Instead, we employ a method known as Factor Analysis on Mixed Data (FAMD), using the the `FAMD()` function in the `FactoMineR` package in R (Husson et al. 2016), which in practice works as a combination of the two methods (see Pagès 2014). Setting aside the external factors, we calculate four FAMDs: one for each class of predictors (internal and stylistic) for each of the two RC types (SRC vs. NSRC). External predictors (Time, Variety, Category), were not subjected to FAMD, to enhance interpretability. We then use the resulting dimensions of the FAMDs as the inputs to our regression models, taking only those dimensions that account for at least 5% of the variance in the data (scree plots for all 4 FAMDs are given in the appendix). The goal is to preserve some degree of interpretability—by keeping internal and stylistics variables separate—while at the same time simplifying our datasets and controlling for complications from data multicollinearity. For both SRC and NSRC models, we included 9 internal and 7 stylistic dimensions as predictors.

The primary drawback of FAMD and related techniques is that the new dimensions cannot always be interpreted easily—they are merely mathematical abstractions derived from complex, multivariate distributions in the data. Finding a reliable influence of "Dimension 3", for example, does not tell us much about the actual linguistic features that affect relativizer choice. Nevertheless, the FAMD output provides measures of the association strength between the original features and each new dimension, and from these associations we can surmise which factors are most influencing the outcome. For continuous variables the FAMD provides the Pearson correlation (*r*) between each variable and dimension. For categorical variables it provides two measures based on one-way ANOVAs: overall explained variance (R^2), and coefficient estimates for each respective category level. Table 1 illustrates how we can interpret the first dimension of the FAMD of the internal features from the subject RC dataset. The results show very strong positive correlations of this dimension with the length of the antecedent NP, the distance from the antecedent head to the RC, and with whether the head and RC are adjacent. Given these measures, it is clear that dimension 1 primarily captures the distance between the antecedent head and the relative clause.

Table 1: Strongest associations of original variables with dimension 1 of internal FAMD of SRC data. Associations of quantitative variables measured as Pearson correlations (*r*). Associations of categorical variables measured as overall explained variance (R^2) and coefficient estimates for each individual category derived from one-way ANOVAs. ('ant' = antecedent)

Numerical		**Categorical**		**Category levels**	
ant head to RC distance	0.907	adjacency	0.434	non-adjacent	2.534
ant Length	0.893	nestedness	0.124	RC = nested	1.147
RC Length	0.140	ant Number	0.044	ant = other num	0.796
		ant POS	0.019	ant = not noun	0.306
		prior Relativizer	0.006	prior = *which*	0.166

ant Givenness	0.003	ant = new	0.080
		ant = Singular	-0.217
		ant = given	-0.080
		prior = *zero*	-0.192
		ant = noun	-0.306
		ant = plural	-0.580
		RC = non-nested	-1.147
		adjacent	-2.534

An advantage of this method is that features which seem conceptually unrelated, yet are highly correlated, will typically be grouped together into the same dimension, e.g. the frequency of subordinating conjunctions and the rate of passivization. This can be especially illuminating in teasing apart stylistic differences, as we discuss below. Techniques such as FAMD thus provide means for reducing the complexity of a huge and unwieldy dataset, with only a minor loss in interpretability. We briefly discuss the associations among FAMD dimensions and original features in the following sections, and provide a complete breakdown of the FAMDs in the supplementary material.

4 Bayesian models of relativizer choice

4.1 General introduction

To test the influence of these multiple factors on relativizer choice, we employ Bayesian multinomial and binomial regression modeling techniques, as opposed to more common regression modeling methods such as those implemented in popular R packages. We believe there are a number of arguments for why linguists might prefer to use Bayesian methods. On a conceptual level, Bayesian inference is more intuitive than the traditional null-hypothesis testing methods used in so-called frequentist statistics. Frequentist methods are so called because they view probability in terms of the relative frequency of an outcome over repeated runs of an experiment. In mathematical terms, the frequentist approach treats the data as random and derives an estimate of a statistic, e.g. a χ^2 value or regression coefficient, along with measure(s) of uncertainty, e.g. the standard error or confidence interval. In other words, frequentist methods, e.g. t or χ^2 tests, provide only a measure of how likely the data would be if the null hypothesis were true.

Bayesian approaches on the other hand, view probability as a measure of a researcher's degree of belief in a hypothesis, i.e. the probability of a hypothesis being true given the observed data. They treat the data as fixed—the facts are what they are—and find the (posterior) probability distribution of some parameter of interest, given the observed data and prior knowledge. Bayesian statistics is thus distinguished by its use of priors. These represent beliefs about the probability of a hypothesis, i.e. the model parameter(s), prior to the observation of any data. The model takes this prior belief into account along with the observed data to arrive at a posterior probability for specific parameter value(s). The posterior probability of a hypothesis depends therefore on both the prior probability and the observed data. Because Bayesian methods necessarily incorporate a researcher's prior beliefs about a hypothesis, they are often criticized as being too "subjective". There are many counter-arguments to this and other criticisms of Bayesian statistics (see, e.g. Lynch 2007:70–73), but the most relevant point is that in practice, the influence of the prior on the posterior tends to be overwhelmed by the influence of the data. In other words, the larger the dataset, the less influence the prior will have on the posterior. Additionally, priors can be specified in such a way as to have as much or as little influence as one would like. For the present study, we used weakly informative priors designed to have minimal influence on the posterior probabilities while permitting (rare) extreme values. Adopting a Bayesian approach allowed us to make use of the most state-of-the-art tools available for addressing our research questions. Furthermore, Bayesian methods offer solutions to the notoriously vexing problems of data sparseness and (near) separation. Finally, in practical terms, our analyses were carried out in R (R Core Team 2015), and at the time of writing there were relatively few well-developed R packages available for multinomial mixed models. The package MCMCglmm (version 2.22, Hadfield 2010) uses Bayesian implementations of a number of generalized linear mixed model types, including binomial and multinomial models, and provides an interface and output similar to that of more familiar modeling packages, e.g. lme4 (Bates, Maechler and Bolker 2015).

4.1.1 Setting up the models. In addition to the use of priors, Bayesian models employ Markov Chain Monte Carlo (MCMC) methods to approximate the posterior distribution. MCMC algorithms generate representative random values from the distribution and then estimate the posterior probabilities from those values, a process often referred to as a 'walk' through the parameter space. The MCMCglmm package uses a combination of common sampling methods, namely Metropolis, Gibbs, and Slice sampling. Both models were fit with 150,000 iterations, with the initial 50,000 iterations removed to correct for initial sampling bias in the chain (so-called 'burn in'). Results sampled every 100th iteration, for a total of 1000 posterior estimates for each model parameter.

Both the NSRC and SRC models included the nine internal and seven stylistic dimensions derived from the factor analyses, as well as the binary predictors of **time** ('1960s' vs. '1990s') and **variety** ('AmE' vs. 'BrE') and their interaction. In terms of the random effects, both models contained by-text and by-category random intercepts, as well as a random intercept for the interaction of corpus and category. The first effect acts as a control for biases from specific authors, while the latter two address potential bias at the level of corpus and text type (genre).

Priors for the random effects and residuals were defined following the suggested structures in the MCMCglmm package course notes (Hadfield 2015). To help control for any potential distortion of the model effects due to (near) separation problems, we specified priors for the fixed effects that are reasonably flat on the probability scale when a logit link is used (see Hadfield 2015:52–55).[3]

Results of Bayesian regression models can be interpreted in a similar fashion to those of frequentist regression models. Whereas frequentist methods derive point estimates of the model coefficients, coefficients in Bayesian models are in fact averages of the posterior distributions over multiple iterations. From these distributions, we can also calculate the 95% Highest

[3] Full details of the model specifications and diagnostics can be found in the supplementary material. Datasets and R code for all analyses can be found online at https://github.com/jasongraf1/relativizers.

Posterior Density (HPD) intervals, also known as 'credible' intervals, which can be used to estimate the probability of a given hypothesis about that predictor's effect. One should bear in mind that Bayesian methods do not use *p*-values for null-hypothesis significance testing. Rather, the posterior distributions can be used directly to estimate the probability of a hypothesis, for example the hypothesis that a regression coefficient is greater than 0. This is measured quite straightforwardly as the proportion of the posterior distribution of that coefficient that falls above 0. Likewise, we can assess directly the probability of the alternative hypothesis, i.e. that the coefficient is less than 0 (see Levshina 2016). All told, if the HPD interval for a given predictor does not contain 0, we can say that the probability of that predictor having some effect (whether positive or negative) is greater than 95%.

4.2 Non-subject extracted RC model

As described above, we first conducted two FAMDs for the NSRC model: one for the internal features, and one for the stylistic features. We included all dimensions that accounted for at least 5% of the variance in their respective FAMD as fixed effects predictors in a Bayesian multinomial regression model. As with binomial logistic regression, multinomial logistic regression estimates the log odds of an outcome given a set of predictor values. Unlike in binomial models however, in the multinomial case we are considering more than two possible outcomes, hence we derive multiple sets of coefficients, reflecting the multiple comparisons the model is making. In general, a model with *n* possible outcomes can make *n* – 1 comparisons. In this case, we have three outcomes—*which*, *zero*, and *that*—and so the model derives two sets of coefficients for each of the two comparisons it makes: *which* vs. *that* and *zero* vs. *that*. That is to say, we treated *that* as the default response in the model.

4.2.1 Model predictions and accuracy. Assessing model fit is not as straightforward with multinomial logistic regression models as it is in binomial logistic regression where there are many statistics for performing model diagnostics, e.g. McFadden's R^2 or Somer's D_{xy}. To measure overall predictive accuracy of our multinomial model, we compare the model's predictions to each observed data point as shown in Table 2 (also referred to as a 'confusion matrix'). We compute the overall accuracy measure as the number of correct predictions across all possible outcomes (the numbers along the diagonal) divided by the total number of observations: (370 + 560 + 2617)/4723 = 0.751. All told, our model correctly predicts the observed relativizer over 75% of the time. This is a notable increase over the baseline accuracy of 60% that we obtain by simply predicting the single most likely relativizer—*zero*—every time. Additionally, we find that for more than three-quarters of our tokens, the model assigns a probability of 0.65 or higher to one of the outcomes. Thus for the majority of cases, the model predicts one outcome with a reasonably high degree of certainty.

Table 2: Non-subject RC model accuracy: Predicted (rows) vs. observed (columns) relativizers

		observed			
		THAT	*WHICH*	*ZERO*	Sum
predicted	*THAT*	370	60	528	958
	WHICH	35	560	337	932
	ZERO	77	139	2617	2833
	Sum	482	759	3482	4723

4.2.2 Internal dimensions. Nine internal dimensions were included as predictors in the NSRC model (see in the Appendix). Table 3 presents the effects of the internal feature dimensions in the NSRC model, which are also represented graphically in Figure 2. We label these dimensions according to the features that correlate most strongly with each, though we emphasize that a given dimension may be associated with many different features. Both show the posterior means and 95% HPD intervals for the effect (in log odds) of each predictor in the model. For the purposes of discussion, we focus on those effects for which the credible interval falls fully above or below 0.

Table 3: Non-subject RC model estimates for internal predictors. Table shows estimated posterior means and 95% HPD intervals for model coefficients, along with the strength of evidence for alternative hypotheses: H_1 = predictor has negative influence on relativizer choice, measured as P(β < 0); H_2 = predictor has positive influence on relativizer choice, measured as P(β > 0).

	posterior mean	0.025%	0.975%	P(β < 0)	P(β > 0)
WHICH:Intercept	0.04	-0.66	0.66	0.44	0.56

WHICH:Int.Dim.1	0.03	-0.05	0.10	0.24	0.76
WHICH:Int.Dim.2	0.12	0.02	0.22	0.02	0.98
WHICH:Int.Dim.3	0.30	0.18	0.42	0.00	1.00
WHICH:Int.Dim.4	0.41	0.26	0.55	0.00	1.00
WHICH:Int.Dim.5	-0.19	-0.33	-0.05	1.00	0.00
WHICH:Int.Dim.6	-0.07	-0.19	0.05	0.87	0.13
WHICH:Int.Dim.7	-0.19	-0.32	-0.07	1.00	0.00
WHICH:Int.Dim.8	-0.10	-0.25	0.04	0.92	0.08
WHICH:Int.Dim.9	0.01	-0.10	0.13	0.41	0.59
ZERO:Intercept	2.21	1.57	2.82	0.00	1.00
ZERO:Int.Dim.1	-0.49	-0.58	-0.40	1.00	0.00
ZERO:Int.Dim.2	-0.37	-0.46	-0.29	1.00	0.00
ZERO:Int.Dim.3	-0.17	-0.27	-0.07	1.00	0.00
ZERO:Int.Dim.4	0.00	-0.10	0.11	0.45	0.55
ZERO:Int.Dim.5	-0.16	-0.26	-0.06	1.00	0.00
ZERO:Int.Dim.6	0.15	0.05	0.25	0.00	1.00
ZERO:Int.Dim.7	-0.07	-0.19	0.04	0.88	0.12
ZERO:Int.Dim.8	-0.64	-0.76	-0.52	1.00	0.00
ZERO:Int.Dim.9	0.19	0.05	0.33	0.00	1.00

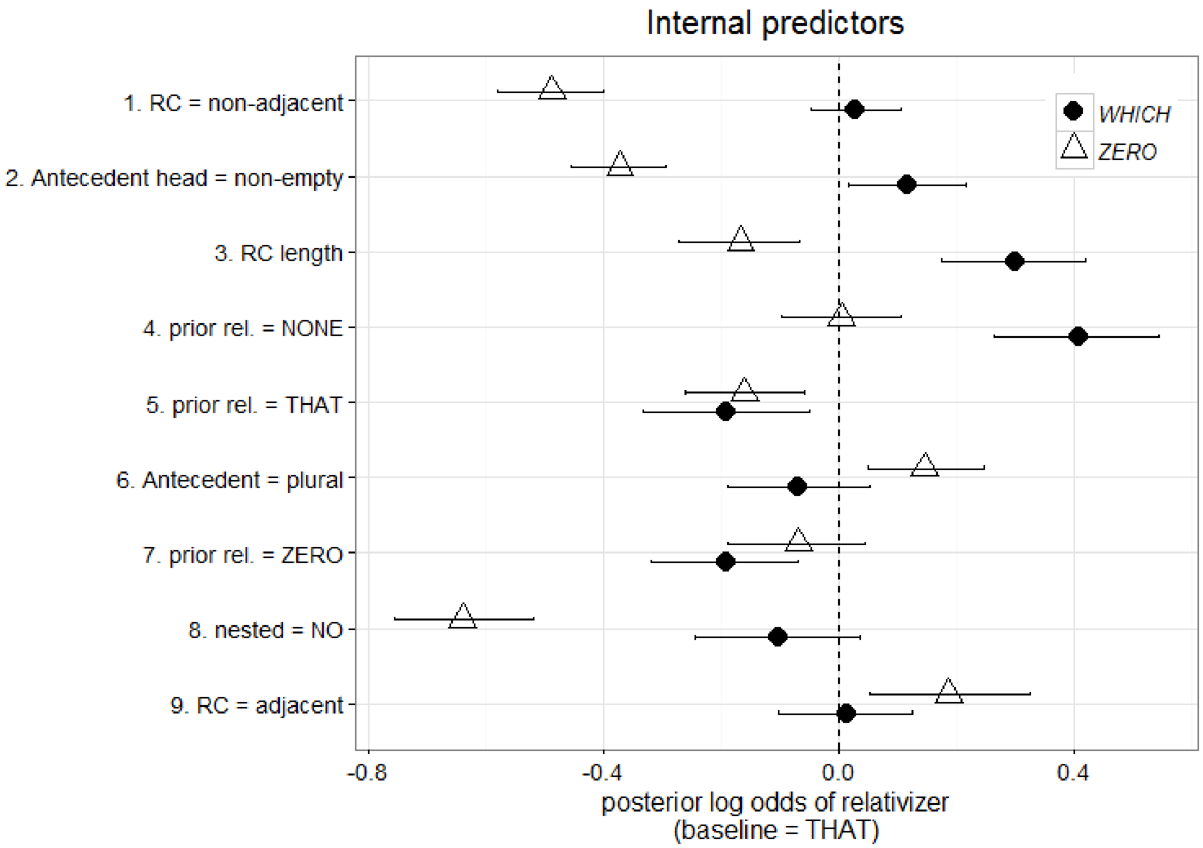


Figure 2: Posterior means and 95% HPD intervals for internal predictors in NSRC model.

The first thing to note is that each of the internal dimensions has some clear role to play in predicting either *which* or *zero* vs. *that*, as indicated by the fact that for each dimension the credible interval of at least one of the outcomes does not cross 0. This is not the case for the stylistic predictors whose effects on relativizer choice are concentrated within only two dimension, as we will see below.

Perhaps the most striking result is the clear contrast between relativizer omission (*zero*) and the use of *which* that we find with dimensions 1-4 and 8. Along with dimension 8, the first three dimensions have a strong negative effect on relativizer omission, while at the same time, dimensions 2, 3, and 4 show a clear positive influence on the choice of *which* over *that*. Upon inspection, we find that dimension 1 is positively correlated with the distance between the antecedent head and the RC (*r* = .86), the overall length of the antecedent (*r* = .73), and the head and RC being non-adjacent (R^2 = .31).[4] In environments where the distance between antecedent and RC is large, e.g. (9), omission is heavily disfavored, though the preference for either of the overt variants is negligible. On the other hand, omission is quite natural in less complex environments, e.g. where the RC immediately follows the antecedent.

(9) *Expo 800, an international trade fair to be held in Dundee from 3 to 6 July 1991* ***which*** *Her Majesty the Queen will open on 3rd July* <F-LOB:E09>

The large influence of dimension 1 can thus be interpreted as a confirmation of the Complexity Principle: more explicit forms are preferred in relatively complex environments.

Dimension 2 is associated primarily with properties of the antecedent, namely part-of-speech, number, definiteness, and length. When the antecedent NP is long, and headed by a definite noun, omission is strongly disfavored and *which* is slightly favored over *that*. Such examples tend to be characteristic of more complex prose style often found in academic (category J) and other non-fiction texts, as illustrated in (10).

(10) *The sort of private interpretations* ***which*** *Wittgenstein was trying to exclude* <FLOB:J51>

The influence of contextual complexity is also partly captured in the effect of dimension 3, which is correlated with longer RCs (r = .59), that are nested (R^2 = .36) and involve a preceding use of *which* (R^2 = .25). In such environments, *zero* is disfavored and *which* is strongly favored over *that* (11). Like with dimension 2, these features tend to occur in more formal writing, so the preference for *which* is to be expected.

(11) *a term [****which*** *is of value to this argument because it hints at the power [****which*** *even conventional linguistic theory attributes to language as the initiator, as well as the descriptor, of social codes]]* <FLOB:J61>

[4] Recall that associations between FAMD dimensions and the original features are measured in terms of Pearson correlations (r) with continuous features, and explained variance (R^2) with categorical features.

The influence of priming is partly captured by dimension 3, and dimensions 4 (R^2 = .86) and 5 (R^2 = .96) are almost exclusively associated with it. Dimensions 3 and 4 are positively correlated with preceding use of *which* or ‘none’ respectively, and so we see a positive preference for *which* over *that*. Dimension 5 also captures the effect of priming, though in this case it is correlated with preceding use of *that*. As we expect, dimension 5 shows that a prior use of *that* has a clear negative influence on the use of either alternative in a subsequent RC. Dimension 7’s clear negative influence on *which* can also be attributed to priming (R^2 = .67), as this dimension is negatively associated with prior use of *which* and positively associated with prior omission.

Two dimensions, 6 and 9, have a positive influence on omission, without favoring either overt alternative. Dimension 6 is most strongly associated with shorter RCs (r = -.32) and plural antecedents (R^2 = .72), while dimension 9 is most strongly associated with longer RCs (r = .33) that are also adjacent to their antecedent heads (R^2 = .43). When considered together, the combined effects of these predictors suggest that the length of the RC plays somewhat less of a role on omission than the degree of integration of the RC. Thus *zero*, but not *which*, is clearly sensitive to features that reflect greater integration, or ‘monoclausality’ (Fox and Thompson 2007:294), of the RC with the noun phrase it modifies.

Finally, we observe a large negative effect of dimension 8 on *zero*. This dimension is associated with longer RCs (r = .47) that involve indefinite antecedents (R^2 = .21), and that are not nested (R^2 = .25) but involve a prior use of *that* (R^2 = .10). This is a bit of a mixed bag. The priming effect is in the expected direction, while the influence of nestedness is not. We take the negative influence of definiteness on *zero* to be a partial confirmation of Wasow et al.’s (2011) findings. The role of RC length however remains unclear.

4.2.3 Stylistic dimensions. Unlike with the internal predictors, the stylistic features seem to have relatively minor influence on the use of NSRCs, as indicated by Table 4. Figure 3 shows the effects of the stylistic predictors that had a reliable impact on relativizer choice in the NSRCs (again, we have provided more informative labels to aid in interpretation). Despite their minor impact, the stylistic effects we do find largely fit with our predictions.

Table 4: Non-subject RC model estimates for Stylistic predictors. Table shows posterior means and 95% HPD intervals for model coefficients, along with the strength of evidence for alternative

hypotheses: H_1 = predictor has negative influence on relativizer choice, measured as P(β < 0); H_2 = predictor has positive influence on relativizer choice, measured as P(β > 0).

	posterior mean	0.025%	0.975%	P(β < 0)	P(β > 0)
WHICH:Style.Dim.1	0.23	0.10	0.35	0	1.00
WHICH:Style.Dim.2	0.10	-0.08	0.28	0.15	0.85
WHICH:Style.Dim.3	-0.13	-0.39	0.13	0.86	0.14
WHICH:Style.Dim.4	-0.06	-0.42	0.31	0.62	0.38
WHICH:Style.Dim.5	-0.04	-0.24	0.16	0.63	0.37
WHICH:Style.Dim.6	-0.04	-0.27	0.19	0.62	0.38
WHICH:Style.Dim.7	0.09	-0.20	0.39	0.25	0.75
ZERO:Style.Dim.1	-0.16	-0.26	-0.07	1.00	0
ZERO:Style.Dim.2	-0.29	-0.46	-0.12	1.00	0
ZERO:Style.Dim.3	0.12	-0.13	0.36	0.14	0.86
ZERO:Style.Dim.4	-0.15	-0.54	0.18	0.80	0.20
ZERO:Style.Dim.5	-0.06	-0.23	0.11	0.76	0.24
ZERO:Style.Dim.6	0.06	-0.12	0.25	0.28	0.72
ZERO:Style.Dim.7	0.15	-0.09	0.38	0.11	0.89

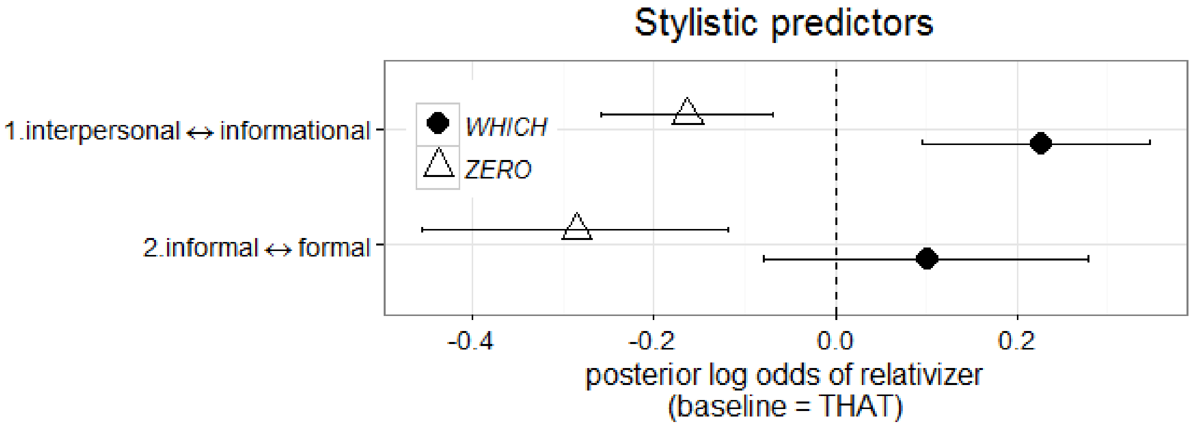


Figure 3: Posterior means and 95% HPD intervals for Stylistic predictors in NSRC model. Log odds values can be interpreted according to the stylistic clines provided: negative values reflect more interpersonal, informal features; positive reflect more informational, formal features.

Stylistic dimension 1 exhibits a negative effect on omission, along with a clear positive influence on the use of *which* over *that*. This dimension is positively correlated with measures of high lexical density, e.g. high noun-verb ratio, nouniness, and mean word length, while it is negatively correlated with features of informal style, e.g. frequent stranding, high numbers of personal pronouns and fictional texts (Table 5). We find that use of an overt relativizer, and in particular *which* is heavily preferred with NSRCs in lexically dense contexts.

Table 5: Association of features with stylistic dimension 1 in FAMD of non-subject relative clause data.

Numerical		**Categorical**		**Category levels**	
noun - verb ratio	0.894	genre	0.680	general prose	0.518
mean Word length	0.886			press	1.080
nouniness	0.749			learned	1.524
passive frequency	0.669			fiction	-3.123
mean Sentence length	0.652				
passive-active ratio	0.511				
type-token ratio	0.271				
subordinating Conj freq.	-0.151				
split infinitive freq.	-0.271				
stranded preposition freq.	-0.449				
personal pronoun freq.	-0.941				

Dimension 2 shows a similar strong negative influence on relativizer omission, but there is weaker evidence for its influence on the choice of *which* vs. *that* ($P(\beta > 0) = .85$). This dimension is positively correlated with measures of high clausal complexity, e.g. high passive-to-active verb ratio, higher average sentence length, and more frequent use of passive verbs and subordinating conjunctions, while simultaneously being negatively correlated with high type-token ratios, and high levels of ‘nouniness’ (Table 6). Furthermore, it is positively correlated with the ‘learned’ genre, and negatively correlated with the 'press' genre. In other words, we find that relativizer omission is significantly less likely in texts that are structurally complex yet low in lexical density, and it would seem such features are characteristic of the academic writing sampled in our corpora.

Table 6: Association of features with stylistic dimension 2 in FAMD of non-subject relative clause data.

Numerical		**Categorical**		**Category levels**	
subordinating Conj freq.	0.627	genre	0.515	learned	2.228
passive-active ratio	0.475			general prose	0.166
passive frequency	0.425			fiction	-0.326
mean Sentence length	0.350			press	-2.068
stranded preposition freq.	0.091				
mean Word length	0.054				
noun-verb ratio	-0.256				
nouniness	-0.446				
type-token ratio	-0.731				

The two factors in Figure 3 thus appear to be capturing two related, but distinct stylistic dimensions. The first places a text along an 'informational' vs. 'interpersonal' spectrum (Biber 1988), where higher values reflect a greater informational focus. The second dimension maps to what we might consider a traditional formal vs. informal division. High values of dimension 2 reflect texts of a more stereotypically formal nature, i.e. they involve greater verbiage and longer, more complex sentence structures. While the two seem conceptually related, they are not coextensive, as, for example, news texts are highly informational yet are known to be moving toward more informal, colloquial styles (Hundt and Mair 1999; Biber 2003). Naturally, interpersonal styles tend to be less formal, hence the strong negative influence on relativizer omission, which is favored in simple, informal style. As dimensions 1 and 2 increase, i.e. lexical density and formality increase respectively, the likelihood of omitting the relativizer decreases. Somewhat surprisingly, we find that *which* is not correlated with formalness per se, but rather its stylistic biases tend toward informationally oriented texts. Why this should be is not clear, though it does suggest that *which* in NSRCs is perhaps more responsive to the demands of reading and producing lexically dense prose, as opposed to serving as a direct marker of heightened formality. Still, while the evidence for an influence of formality on *which* is not as strong as it is for information density, we note that there is nonetheless an 85% likelihood that these features have some influence in the expected direction.

4.3 Subject extracted RC model

Like with the NSRC case, we first conducted two separate FAMDs for the SRC model for the internal features and for the stylistic features, again including only those dimensions that accounted for at least 5% of the variance in their respective FAMD. Again, nine internal and seven stylistic dimensions, along with the external predictors, were included as fixed effects predictors in a Bayesian binomial regression model. The predicted outcome in the binomial model was *which*.

4.3.1 Model predictions and accuracy. Overall, our SRC model predicted relativizers with a high degree of accuracy (C = 0.93, D_{xy} = 0.85). Comparison of the model's predictions to each observed data point is shown in Table 7, from which it can be seen that the model predicts 84.4% of the instances correctly. This is a significant increase over the baseline accuracy of 56.6% that we obtain by predicting that every time ($p_{\text{binom}} \approx 0$).

Table 7: Subject RC model accuracy: Predicted (rows) vs. observed (columns) relativizers

		observed		
		THAT	*WHICH*	Sum
predicted	*THAT*	5032	787	5819
	WHICH	814	3652	4466
	Sum	5846	4439	10285

4.3.2 Internal dimensions. Main effects of the internal feature dimensions in the subject RC model are presented in Table 8. We have strong evidence for a reliable influence of six out of the nine predictors on the choice of relativizer, with all but dimension 2 exhibiting a positive influence on the likelihood of using which instead of that. These six dimensions are shown graphically in Figure 4, labelled according to their strongest respective associations with the original features.

Table 8: Subject RC model estimates for internal predictors. Table shows posterior means and 95% HPD intervals for model coefficients, along with the strength of evidence for alternative hypotheses: H_1 = predictor has negative influence on relativizer choice, measured as P(β < 0); H_2 = predictor has positive influence on relativizer choice, measured as P(β > 0).

	posterior mean	0.025%	0.975%	P(β < 0)	P(β > 0)
(Intercept)	-0.53	-1.41	0.30	0.88	0.12
Int.Dim.1	0.11	0.06	0.15	0.00	1.00
Int.Dim.2	-0.23	-0.29	-0.17	1.00	0.00
Int.Dim.3	0.35	0.28	0.42	0.00	1.00
Int.Dim.4	0.21	0.14	0.28	0.00	1.00
Int.Dim.5	0.05	-0.02	0.11	0.08	0.92
Int.Dim.6	0.05	-0.02	0.12	0.10	0.90
Int.Dim.7	0.14	0.07	0.22	0.00	1.00
Int.Dim.8	-0.03	-0.10	0.05	0.79	0.21
Int.Dim.9	0.10	0.02	0.18	0.01	0.99

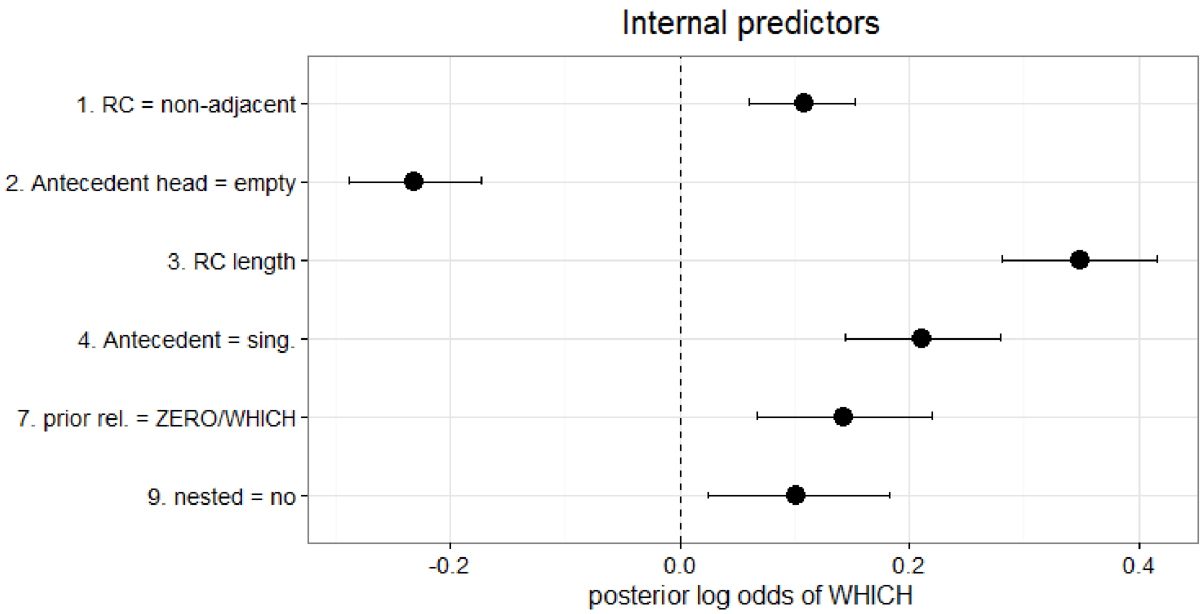


Figure 4: Posterior means and 95% HPD intervals for internal predictors in SRC model.

As the NSRC analysis, the first dimension of the subject-extracted RC FAMD expresses the distance between the antecedent head and the RC. This is clear from dimension 1's high correlation with the head-to-RC-distance (r = .91), the total length of the antecedent NP (r = .89), and the binary factor **adjacency** (R^2 = .44). When the distance from the antecedent head to the RC is large, *which* is the preferred variant in SRCs. Assuming that *which* is the more marked, and hence explicit variant, this result conforms to the Complexity Principle.

Dimension 2 in the FAMD also captures features of the antecedent, namely number (R^2 = .70) and part-of -speech (R^2 = .69), as it does in the NSRC analysis, though in a slightly different fashion. Unlike in the NSRC model, which shows a positive influence of dimension 2 on the likelihood of *which*, this dimension in the subject RC model has a strong negative influence on the likelihood of *which*. This is due to the fact that dimension 2 in the SRC model is positively correlated mostly with RCs whose antecedents are **not** nouns, and were coded as neither singular nor plural (i.e. as 'other'). These are cases where the antecedent is a lone adjective or quantifier, as in (12) for example. Such contexts strongly favor the use of *that* over *which*.

(12) *The most **that** was accomplished was adding Mrs. Beige's tray to the dish pile…*
<Brown:R02>

Fox and Thompson (2007) refer to such antecedents as ‘Empty Head’ NPs, and they show that ‘empty’ antecedents tend to favor relativizer omission in NSRCs. Naturally, this does not apply to SRCs, since the *zero* option is unavailable, however it is very likely that the underlying forces driving this tendency in the NSRC context are also at work in SRC contexts. As with the preference for *zero* with such antecedents in NSRCs, we suggest that the preference for the least marked/explicit variant *that* in SRCs is due to the RCs higher degree of integration with the NP, which, arguably, also correlates with the probability of an upcoming RC given an empty antecedent. This use of the unmarked option in highly predictable and/or integrated SRCs is compatible with a variety of approaches to relativizer omission in NSRCs (e.g. Temperley 2003; Fox and Thompson 2007; Jaeger 2010).

Dimensions 3, 4, and 7 are all associated with priming effects to varying degrees, and all three exhibit positive influence on the likelihood of *which*. Dimensions 4 and 7 are both positively correlated with prior instances of relativizer omission (*zero*) and negatively correlated with prior use of *that* (13), while dimension 3 is positively associated with a prior use of *which*, and negatively associated with prior omission (14). Dimension 3 also exhibits a strong positive correlation with RC length (r = .72), again illustrating the preference for *which* in more complex environments. Note that this result contrasts with what was found in NSRCs, where use of *which* was only very weakly correlated with RC length.

(13) *And once elk have found a good source of food, they’ll want to conserve the body heat* ***Ø*** *it produces by finding a place to hide and bed* **that** *will minimize heat loss*. <Frown:E10>

(14) *Children should investigate those materials* ***which*** *conduct electricity and those* ***which*** *do not.* <FLOB:H03>

Finally, dimension 9 is also positively correlated with RC length (r = .50), as well as nestedness (R^2 = .31).

4.3.3 Stylistic dimensions. In Table 9 we present the results for all seven stylistic dimensions in the SRC model. Figure 5 shows the two stylistic dimensions that had a reliable influence (P(β) = 100%) on the use of which in the SRC model. We find that the influences of stylistic features on the SRC relativizers are captured primarily by these first two dimensions, just as they in the NSRC model.

Table 9: Subject RC model estimates for stylistic predictors. Table shows posterior means and 95% HPD intervals for model coefficients, along with the strength of evidence for alternative hypotheses: H_1 = predictor has negative influence on relativizer choice, measured as P(β < 0); H_2 = predictor has positive influence on relativizer choice, measured as P(β > 0).

	posterior mean	0.025%	0.975%	P(β < 0)	P(β > 0)
Style.Dim.1	0.35	0.26	0.44	0.00	1.00
Style.Dim.2	0.33	0.17	0.48	0.00	1.00
Style.Dim.3	0.06	-0.13	0.25	0.25	0.75
Style.Dim.4	-0.05	-0.29	0.19	0.66	0.34
Style.Dim.5	-0.03	-0.29	0.23	0.61	0.39
Style.Dim.6	-0.04	-0.29	0.20	0.63	0.37
Style.Dim.7	0.02	-0.17	0.23	0.42	0.58

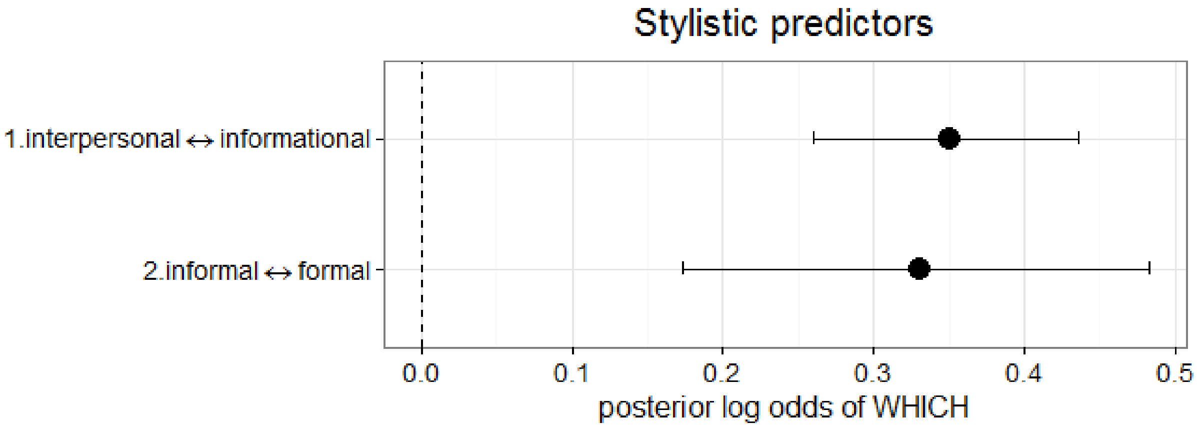


Figure 5: Posterior means and 95% HPD intervals for Stylistic predictors in NSRC model. Log odds values can be interpreted according to the stylistic clines provided: negative values reflect more interpersonal, informal features; positive reflect more informational, formal features.

The first two dimensions again distinguish between highly informational texts (dimension 1) and texts that exhibit features indicative of greater formality (dimension 2). We present the associations of those dimensions and the original features in Tables 11 and 12.

Table 10: Associations of original features with stylistic dimension 1 in FAMD of subject relative clause data.

Numerical		**Categorical**		**Category levels**	
noun-verb ratio	0.855	genre	0.605	learned	1.714
mean Word length	0.836			press	0.869
nouniness	0.690			general prose	0.518
passive frequency	0.644			fiction	-3.216
passive-active ratio	0.542				
mean Sentence length	0.482				
split infinitive freq.	-0.190				
subordinating Conj freq.	-0.273				
stranded preposition freq.	-0.377				
personal pronoun freq.	-0.926				

Dimension 1 clearly falls along an informational vs. interpersonal cline, as reflected by its very strong positive correlations with markers of informationally rich texts[5], e.g. nouniness and mean word length, in tandem with its strong negative correlations with use of features characteristic of colloquial, interpersonal style, e.g. personal pronouns and preposition stranding (Biber 1988). Furthermore, we find that **genre** explains a great deal of the variance in dimension 1, being positively associated with non-fiction, news, and especially academic ('learned') texts, while

[5] Note this is not simply a matter of greater lexical density, as type-token ratio does not correlate at all with dimension 1.

being negatively associated with fiction texts. The results of the model show that the more "informational" a text, the greater the probability of *which* over *that*.

Turning to the associations of dimension 2, we find this dimension to be a measure of a text's compactness. Dimension 2 is positively correlated with features reflective of greater sentential complexity, e.g. use of subordinating conjuncts, passivization, and greater sentence length, while also being negatively correlated with measures of lexical density, e.g. nouniness and especially TTR. The explanatory power of **genre** for this dimension is again quite high, though in this case the primary contrast is between academic texts and news texts. Newspaper texts, being especially compact, are naturally negatively associated with this dimension. In short, dimension 2 is a measure of prolixity, and to the extent that increased verbiage is seen to be characteristic of more formal style, we interpret this dimension as a metric of formality. This interpretation is also congruent with the demonstrated trend toward more informal style in newspaper texts in the 20th century (Hundt and Mair 1999; Biber 2003). As we predict, the model shows that *which* becomes more likely as the formality of a text increases.

Table 11: Associations of original features with stylistic dimension 2 in FAMD of subject relative clause data.

Numerical		**Categorical**		**Category levels**	
subordinating Conj freq.	0.569	genre	0.504	learned	1.816
passive-active ratio	0.516			general prose	0.190
passive frequency	0.472			fiction	-0.053
mean Sentence length	0.318			press	-1.953
stranded preposition freq.	0.219				
mean Word length	-0.024				
noun-verb ratio	-0.321				
nouniness	-0.542				
type-token ratio	-0.801				

5 Discussion

Results of our analysis are somewhat more complex than in typical studies of this nature (e.g. Leech et al. (2009); Hinrichs et al. 2015), so let us briefly take stock of what we have found. On the one hand, we find patterns that largely confirm findings from previous studies, e.g. the decreasing preference for *which* over time in American English (Leech et al. 2009; Hinrichs et al. 2015), and the role of complexity and priming in predicting relativizer choice (Szmrecsanyi 2006; Rohdenburg 1996; Rohdenburg 2014; Hinrichs et al. 2015). On the other hand, we find new evidence for variation in the strength of certain predictors' effects depending on the function of the relative pronoun. In particular, we find that complexity—as measured in the distance between the antecedent and the RC—exerts a weaker influence on the choice of *which* in NSRCs than in SRCs, as do stylistic features associated with greater sentential complexity. Returning then to the main question we set out with—do the same factors affect relativizer choice in both SRCs and NSRCs?—we find that the answer is "yes, but not necessarily to the same degree". With this in mind, we frame the discussion of overt relativizer variation against the backdrop of our findings vis-à-vis relativizer omission. We take omission as our starting point simply because we know much more about the internal, cognitive factors motivating the *zero* option than we do about similar factors motivating the choice of *which* vs. *that*. This paper is one step toward closing this gap in our understanding.

5.1 Relativizer omission

Based on much previous work, we expected the use of overt vs. *zero* markers to be influenced by a number of internal and external factors. Internal factors included the relative complexity of the RC context, as measured by RC length and distance between antecedent and RC, prior use of specific relativizers (priming), and lexical properties of the antecedent heads, i.e. empty vs. non-empty heads. In addition to internal factors, we predicted that *zero* should be favored in more informal texts, all else being equal, but expectations regarding its variation across time and region were less clear.

Turning first to the internal factors, we find that relativizer omission in NSRCs behaved mostly as we expected, though with some slight new twists. First and foremost, there is strong evidence that the greater distance between the antecedent head and the RC the higher the likelihood of *that* over *zero*. While this distance is one of the metrics of 'cognitive complexity' used to support complexity-based approaches (e.g. Rohdenburg 1996; Fox and Thompson 2007), we believe this finding is compatible with a number of possible explanations for relativizer omission, including ambiguity avoidance (Temperley 2003) and probability (Levy and Jaeger 2007; Wasow, Jaeger and Orr 2011) based accounts. On a similar note, we find that RC length also plays a role in predicting omission, with *zero* being more likely when the RC is short. Both RC length and antecedent-RC adjacency have been subsumed under the notion of 'complexity' by a number of authors (e.g. Quirk 1957; Fox and Thompson 2007; Rohdenburg 2014), however it is not yet clear how ambiguity avoidance or probabilistic approaches might account for the effects of RC length. We return to this issue below.

A second major factor that emerged from our model(s) is that of syntactic priming, where prior use of an overt relativizer decreased the likelihood of *zero* in the target NSRC. This effect was captured in the influence of a number of dimensions in the FAMDs. We find some evidence that prior omission decreases the odds of subsequent omission (dimension 7), but the evidence is not as strong as for the priming effect of overt relativizers. We also observe a positive influence of 'empty' antecedent heads on omission (dimension 2), as well as an influence of number and, to a lesser extent, definiteness (dimension 6). In our data, *zero* is more likely than *that* when the antecedent is plural and/or definite, or one of a number of lexically non-specific heads such as *all*, (*some/every/any*)*thing*, and *one*. We hypothesize that the effect of such 'empty' heads (and possibly number) is a reflection of their high degree of entrenchment (or predictability) as RC antecedents (see e.g. Wasow et al. 2011; Wiechmann 2015).

Finally, in terms of stylistic patterns, we find that *zero* is overall favored in the informal and/or interpersonal prose typical of fiction and newspaper texts. We also find new evidence for a slight decline in relativizer omission over time (cf. Hinrichs et al. 2015), and this decline appears to be slightly more pronounced in America than in Britain. Why this should be is not clear at this time, and we hold off speculating until this trend can be more fully verified.

5.2 Variation in that vs. which

The primary goal of this study however was not to explore relativizer omission itself, but rather to explore variation in the choice between overt variants across subject and non-subject RCs, while simultaneously taking account of the *zero* option. Nonetheless, the influence of various factors on *zero* is congruent with their effects on the choice between *which* and *that*. First, we find clear evidence for a priming effect operating in the same direction in both SRCs and NSRCs. When the preceding RC is introduced by *that* or *zero*, the likelihood of *which* with a following RC decreases noticeably, and this effect does not appear to operate differently in SRC or NSRC contexts. We further find that empty antecedents disfavor the use of *which* and favor the use of the unmarked variant, i.e. *that* in SRCs and *zero* in NSRCs. This effect accords with the Complexity Principle, as *which* is the more explicit/marked variant based on various phonological, semantic, and distributional properties; however, the effect is also compatible with probabilistic accounts, assuming that such heads often occur with relative clauses (Wasow, Jaeger and Orr 2011). For example, UID models (e.g. Jaeger 2011) predict that when multiple syntactic options are available, users will choose the option that optimizes information flow, thus the more reduced form is favored when the RC is predictable. Conversely, we could say that the more explicit variant is preferred when the RC is less predictable. If RCs tend to be more common following such empty-headed forms, probability-based accounts like the UID model would predict that *which,* the more explicit variant, should be less likely with RCs involving empty antecedents. Things become more interesting when we consider the effects of more direct complexity measures, namely RC length and antecedent-RC adjacency.

As complexity-based accounts predict, we find that longer RCs do indeed favor the use of *which*, regardless of relativizer function. As with omission, it is not immediately clear how to explain this effect in terms of incremental processing models, such as the UID model. It is possible that, since longer structures are naturally less likely than shorter ones, the effect of RC length could nonetheless be a reflection of the tendency for less predictable RCs to take more marked relativizers, possibly as a means of buying processing time for the language user (or alternatively, providing a more informative cue of an upcoming RC to the listener). At present however, we are unaware of any work demonstrating a correlation between the likelihood of an RC (given the preceding information) and length or structural complexity. This possibility could be tested directly by, for example, looking at the likelihood of RCs of a particular length or structural shape given the preceding material.[6]

One factor for which we do find differences across the two RC functions is the distance between the antecedent and the RC (dimension 1 in both FAMDs). In SRCs, greater distance between antecedent and RC increases the odds of *which* over *that* substantially; however, the same is not true with NSRCs. The effect trends in the same direction, but the evidence for NSRCs is not nearly as strong. While the general effect of antecedent-RC adjacency in this direction is expected, the contrast between SRCs and NSRCs is not necessarily predicted by any account, as far as we can tell. The discourse-functional specialization of SRCs and NSRCs has been noted by some (e.g. Fox and Thompson 2007; Wiechmann 2015), but mainly in the context of omission. How this might pertain to choice among overt relativizers remains an open question.

A second difference between SRC and NSRC relativizers that emerged from our study lies on the stylistic plane. The FAMDs of predictors in both SRC and NSRC datasets identify two related, though conceptually distinct, stylistic clines: an informational vs. interpersonal cline, and a formal vs. informal cline (see also Biber 1988). In both functions the use of *which* is favored in texts marked by high degrees of informational richness, as reflected in greater degrees of nouniness and average word length, and disfavored in more interpersonal prose, especially fiction. By contrast, we find that the more direct association between *which* and common markers of formality (e.g. more passivization, longer average sentences, lower average TTR) is much stronger in SRCs than in NSRCs. This suggests that the use of *which* as a marker of formality is somewhat more flexible when it introduces a subject rather than non-subject RC. We stress though that the evidence for an influence of formality on non-subject relativizers is still quite strong, with there being an 85% probability that use of *which* over *that* increases with a text's degree of formality. This is not at all a "null" result, in other words. We suggest that there simply may not be enough information in our NSRC data to determine whether *which* has an effect with

[6] We thank an anonymous reviewer for this suggestion.

the same degree of confidence as with the SRC data, however we do find some compelling evidence in the expected direction.

Finally, we note that the effects of our stylistic features accord with the findings of Hinrichs et al. (2015), who show that frequent use of passive verbs reliably predicts greater usage of *which* over *that*. Given the strong positive correlations of stylistic dimensions 1 and 2 with frequent passivization, the present findings can be seen as partial confirmation of Hinrichs et al's study.

5.3 'Complexity', processing, and predictability

But why should we find influences of 'Complexity' on the choice between overt variants, and why should there be a difference in this effect across subject and non-subject RCs? If, as we propose, *which* is the more explicit variant, the Complexity Principle predicts that it should be preferred in more complex environments, and this is generally what we find. While there is little doubt that the Complexity Principle makes a reliable predictions about a number of phenomena (Rohdenburg 2014), it nonetheless remains an open question as to what exactly 'Complexity' is in this view. In practice, Complexity is often treated as a composite of features thought to affect processing difficulty, including number of words/nodes in a phrase, pronominality, and sententiality (e.g. Fox and Thompson 2007; Berlage 2014). However, it not always made clear how these features are expected to affect processing. Many have argued that complexity-related effects, e.g. the principle of End Weight, are driven by the desire to minimize dependencies between grammatical elements (e.g. a head word and its complement), where users prefer word orders with shorter dependencies (e.g. Hawkins 2004). The dependency minimization strategy is motivated by general cognitive limitations: shorter dependencies are less taxing on working memory. The dependency minimization approach has been argued to explain numerous word order effects, and is compatible with findings from relativizer omission, but the predictions of such accounts on the choice of *which* over *that* are not as clear. Nor is it obvious from this point of view how or why properties of the RC itself should affect the choice of any variant, *zero* or overt, in the way that they do.

Alternative processing accounts do make predictions about the influence of the RC properties, as well as other processing effects that are not explained by dependency minimization models, for instance priming. These accounts hold that production is heavily shaped by the accessibility of linguistic units, where more accessible lexical items or constructions are more likely to be produced before less accessible ones. Accessibility in these approaches is usually defined as the ease of retrieval from memory (MacDonald 2013; Wiechmann 2015). With regard to relativization strategies, accessibility accounts make a number of predictions that are borne out in the data. First, they predict that previous uses of specific relativizers render those relativizers more accessible, hence they should prime subsequent use of those same variants. Indeed that is what we find. Second, we expect properties of the RC to have some role, as language users tend to produce overt variants when the upcoming material is less accessible, and hence unavailable for production (Jaeger and Wasow 2006). Alternatively under an entrenchment account, properties of the RC are obviously properties of the construction as a whole, thus relevant to its degree of entrenchment, where entrenchment is a direct function of frequency (Wiechmann 2015). Finally, though we acknowledge that the relationship between length and accessibility is not direct (see Jaeger 2011:165–166), there is a clear tendency for length to be inversely correlated with accessibility (Hawkins 2004; MacDonald 2013). Thus we find that the length of the RC does indeed influence the choice of variant, with *zero* being less likely before longer RCs.

Multiple factors such as imageability, conceptual (e.g. animacy) and/or contextual (e.g. givenness) salience, and of course frequency may contribute to the accessibility or "easiness" (MacDonald 2013) of a linguistic element, in this case the particular form of a functional lexical item. As mentioned in section 2.2, frequency effects are sometimes couched within an information-theoretic framework, which views the effects of surface and structural level probabilities and co-occurrence frequencies as the result of biases in production aimed at maintaining a constant rate of (Shannon) information transmission (e.g. Levy 2007; Jaeger 2011). In contexts where an RC structure is more predictable, i.e. more likely given the preceding material, users will opt for the less redundant, i.e. more reduced variant. We suggest the predictability-centered approach also provides a tentative explanation for the influence of antecedent-RC adjacency: the greater the distance between the antecedent and RC, the less likely a relative clause becomes.[7] It should be noted that although recent work has tended to focus on structural and surface level probabilities, the predictability of a given element is no doubt sensitive to other factors such as semantic biases or contextual cues (see also Wasow et al. 2011). Thus, predictability, as derived from distributional patterns in the input, can be unified with accessibility under a Production-Distribution-Comprehension model which argues that

> structure choices in production, at least some of which are determined by production-specific mechanisms, create robust distributional patterns in the language, which are learned over time by comprehenders who are exposed to this input. These distributional patterns then become the probabilistic constraints that guide the comprehension process in a constraint-based system. (Gennari and MacDonald 2009:2)

Taking inspiration from the Complexity Principle, but working from a probabilistic framework, we propose the following extension of the Predictability Hypothesis[8] of Wasow et al. (2011:5):

(15) The more predictable a relative clause is given the preceding material, the more likely speakers/writers are to use the most reduced, i.e. least explicit, relativizer option.

It must be stressed here that the notion of what is 'most reduced' must be restricted to a given variable context. That is to say, we do not claim that high predictability should always lead to maximal reduction, i.e. omission, in all cases. A full account of syntactic variation must leave room for the grammar to constrain variation in multiple ways. In absolute terms, the complete

[7] We also follow Wasow et al. (2011) in observing that the notion of predictability has advantages over approaches such as Fox and Thompson's (2007) notion of 'monoclausality', in that it provides a definable—and psychologically grounded—measure with which to gauge the degree of integration between the RC and the NP it modifies. At the same time, we also note that a more complete probabilistic model should be able account for highly schematized structures, as in Wiechmann's (2015:191-195) discussion of the discourse-functionally entrenched construction *it's something that X.* What such a model might look like is a question we leave for the future.

[8] While Wasow et al. restrict their hypothesis to the domain of omission in NSRCs, our generalization in (15) is entirely in line with the spirit of their proposal.

absence of a form is of course the maximal case of reduction, however not all contexts allow this degree of reduction. In relative terms then, we might tentatively define the most reduced form as the most frequent and/or least explicit variant *of those available in the variable context*. As we suggest for *which*, the explicitness of a given variant may be dependent upon any number of relevant stylistic, structural, or semantic dimensions. In the case of NSRCs, the maximally reduced variant is *zero*—nothing at all—but in the case of SRCs, StE grammar simply does not allow the complete omission of the relativizer. But there is less explicit, i.e. more frequent, more semantically ambiguous, and less phonologically marked variant: *that*.

6 Conclusion

The purpose of this study was to investigate the extent to which the factors that affect choice among relativizers in subject-extracted relative clauses also affect relativizer choice in non-subject-extracted relative clauses. Results show that variation in both types of relative clause contexts is shaped largely by the same forces, though we find tentative evidence that the strength of certain factors may vary between subject and non-subject RCs. These factors include both psycholinguistically driven effects, e.g. cognitive ‘complexity’, as well as stylistic differences in the preference for *which* in more formal texts. Exactly why these factors should vary in the way they do is not entirely clear, however we believe a probabilistic approach to understanding relativizer variation situated within a unified Production-Distribution-Comprehension model provides a plausible framework for explaining our results.

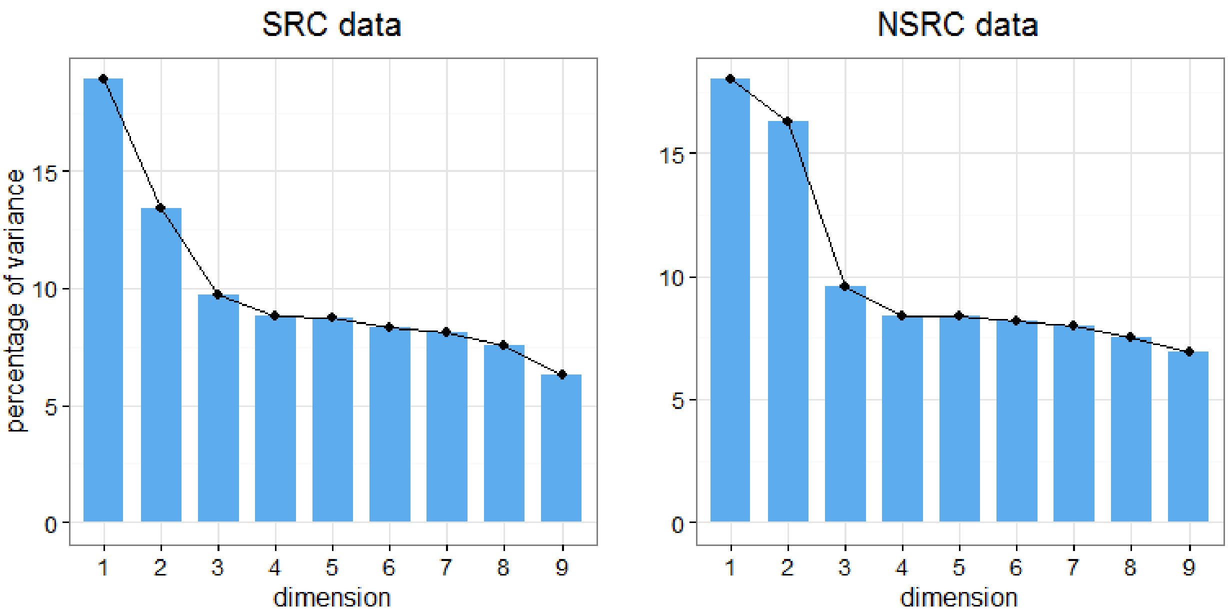
On the methodological plane, we used two statistical techniques that are relatively novel and somewhat unorthodox in variationist studies of this kind. The high degree of redundancy in natural language—where any single feature may be predictable from many others—motivated our use of dimension reduction techniques to help identify orthogonal dimensions of variability in our data. While such techniques can potentially obfuscate effects of specific linguistic features, our analyses nonetheless identified easily interpretable and linguistically coherent dimensions which have strong predictive influence on relativizer choice. In addition, we advocate the use of Bayesian regression analysis in variationist studies, for both practical and epistemological reasons. Most importantly, Bayesian methods allow us to examine the strength of evidence for one or multiple hypotheses via direct measures of their respective probabilities of being true given the data. Such results provide much richer information for generating future hypotheses.

To that point, we believe our findings suggest a number of avenues for future research. As a start, a more complete accounting of the full range of contextual cues would be desirable. Features of the main clause and discourse function of the RC were not included in the analysis, yet there is some evidence that they play a role in relativizer choice (e.g. Fox and Thompson 2007; Wiechmann 2015). At the same time, processing related explanations for relativizer choice patterns have become widely popular; however, our results provide mainly indirect evidence for such accounts, as testing such accounts was not our primary focus. We believe a more precise investigation of RC predictability could potentially offer a number of insights into these and others' findings. Finally, we remain curious to what extent the patterns we find in written English parallel those of spoken (standard) English, and to what extent patterns in production data from written corpora are reflected in aspects of language comprehension.

Appendix

Table 12: Condition numbers (κ) for predictor groups entered into the factor analyses. Kappa scores represent the degree of multicollinearity in the design matrix of a hypothetical regression model containing those predictors. Values above 30 are generally considered to be potentially harmful (Baayen 2008; Belsley et al. 2004).

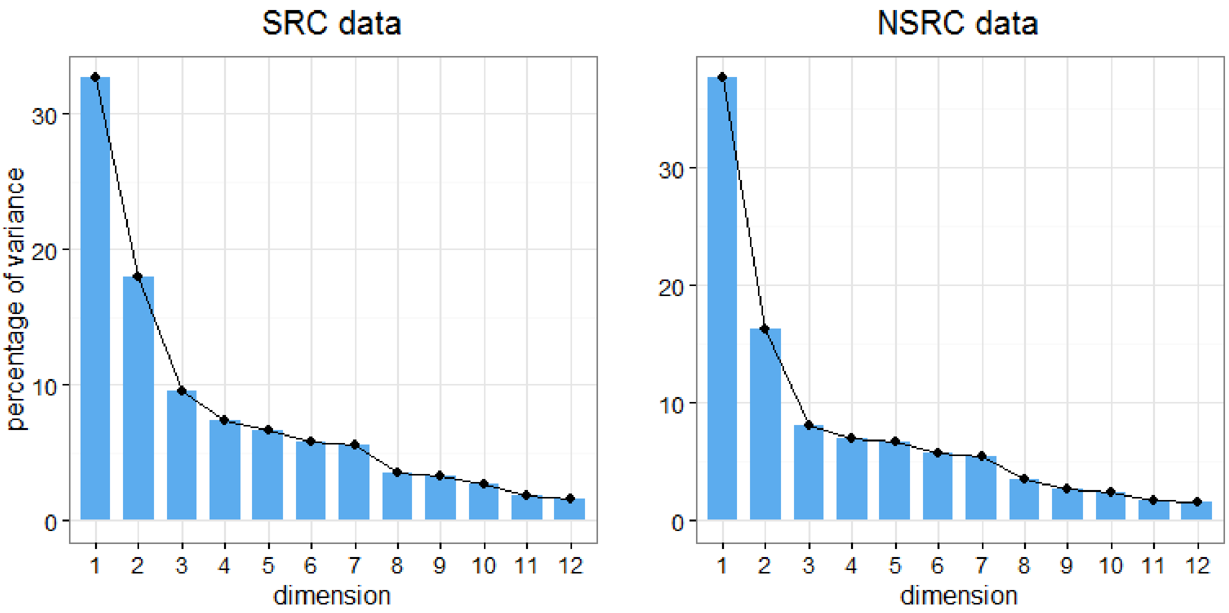
	Predictor groups	
	Internal	Stylistic
Subject RCs	36.2	108.6
Non-subject RCs	44.3	130.1



SRC data

NSRC data

Figure A.1: Scree plots of FAMDs for internal predictors in SRC and NSRC datasets. Plots display the percentage of total variance in the data each dimension accounts for.



SRC data

NSRC data

Figure A.2: Scree plots of FAMDs for stylistic predictors in SRC and NSRC datasets. Plots display the percentage of total variance in the data each dimension accounts for.

External predictors in NSRC model

We find suggestive evidence for a main effect of variety with both *which* and *zero*, with *which* being slightly favored ($P(\beta > 0) = .92$) and *that* slightly disfavored ($P(\beta < 0) = .91$) by British writers as compared to American authors. We also find noticeable decreases in both relativizers over time, and that the use of *which* over time has decreased substantially more so in American English than in British English. There is also some evidence for a further decrease of *zero* in American English vs. British English, though the case is not as strong as for *which*. Thus, while the *zero* option remains the more frequent overall, it is losing ground to *that*, as reflected in the (slight) downward trend of zero in both varieties in Figure 6. This is striking given that the raw frequencies would suggest a trend in the opposite direction (see Figure 1).

Table 13: Non-subject RC model estimates for external predictors. Table shows posterior means and 95% HPD intervals for model coefficients, along with the strength of evidence for alternative hypotheses: H_1 = predictor has negative influence on relativizer choice, measured as $P(\beta < 0)$; H_2 = predictor has positive influence on relativizer choice, measured as $P(\beta > 0)$.

	posterior mean	0.025%	0.975%	P(β < 0)	P(β > 0)
WHICH:Intercept	0.04	-0.66	0.66	0.44	0.56
WHICH:varietyBrE	0.52	-0.17	1.27	0.08	0.92
WHICH:time21991	-2.43	-3.14	-1.66	1.00	0
WHICH:varietyBrE:time21991	1.55	0.44	2.61	0.01	0.99
ZERO:Intercept	2.21	1.57	2.82	0	1.00
ZERO:varietyBrE	-0.48	-1.18	0.26	0.91	0.09
ZERO:time21991	-0.84	-1.48	-0.14	1.00	0
ZERO:varietyBrE:time21991	0.62	-0.34	1.55	0.11	0.89

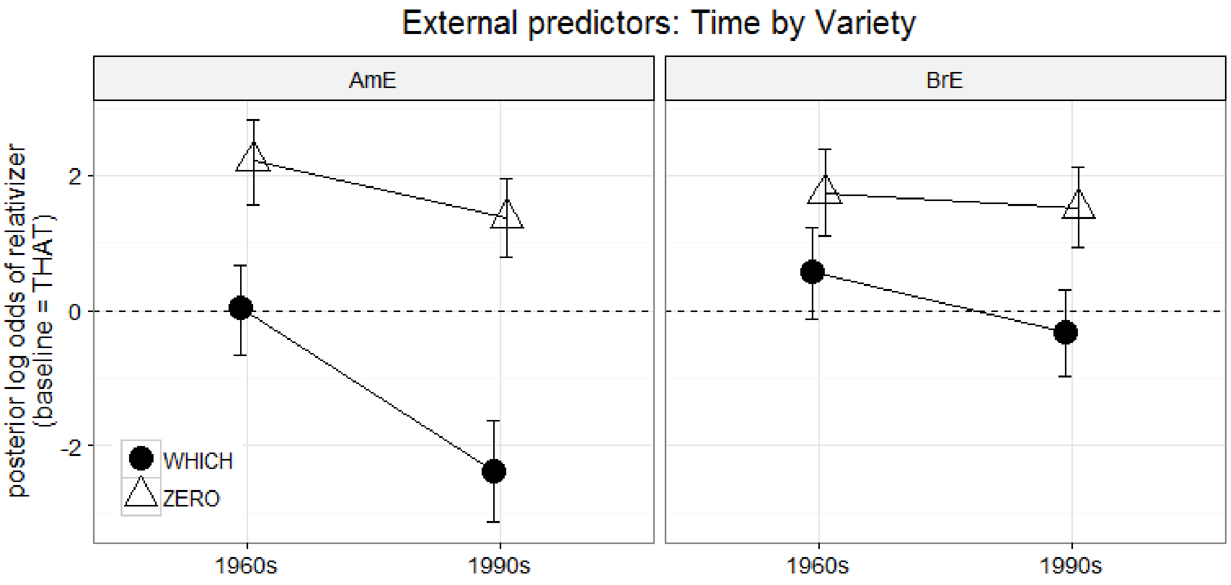


Figure 6: Posterior means and HPD intervals for NSRC relativizers across Time and Variety

External predictors in SRC model

Results of our external predictors largely accord with what we expected from previous research. We find a clear overall preference for *which* in British English vs. American English, as well as a general trend away from *which* over time. We further find that this diachronic trend has been much stronger in the U.S. than in Britain (Figure 7). The findings from the SRC data thus parallel those from the NSRC data.

Table 14: Subject RC model estimates for external predictors. Table shows posterior means and 95% HPD intervals for model coefficients, along with the strength of evidence for alternative hypotheses: H_1 = predictor has negative influence on relativizer choice, measured as P(β < 0); H_2 = predictor has positive influence on relativizer choice, measured as P(β > 0).

	posterior mean	0.025%	0.975%	P(β < 0)	P(β > 0)
(Intercept)	-0.53	-1.41	0.30	0.88	0.12
varietyBrE	1.19	0.29	2.13	0.01	0.99
time21991	-2.90	-3.90	-1.90	1.00	0.00
varietyBrE:time21991	2.74	1.51	3.87	0.00	1.00

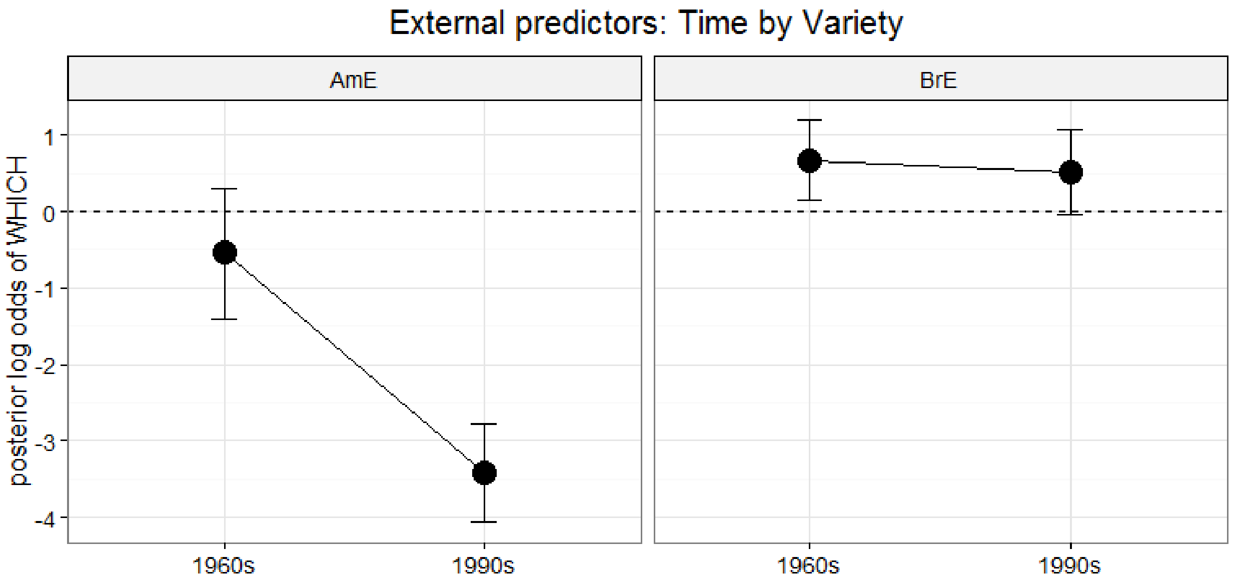


Figure 7: Posterior means and HPD intervals for NSRC relativizers across Time and Variety

Subgenre random adjustments in regression models

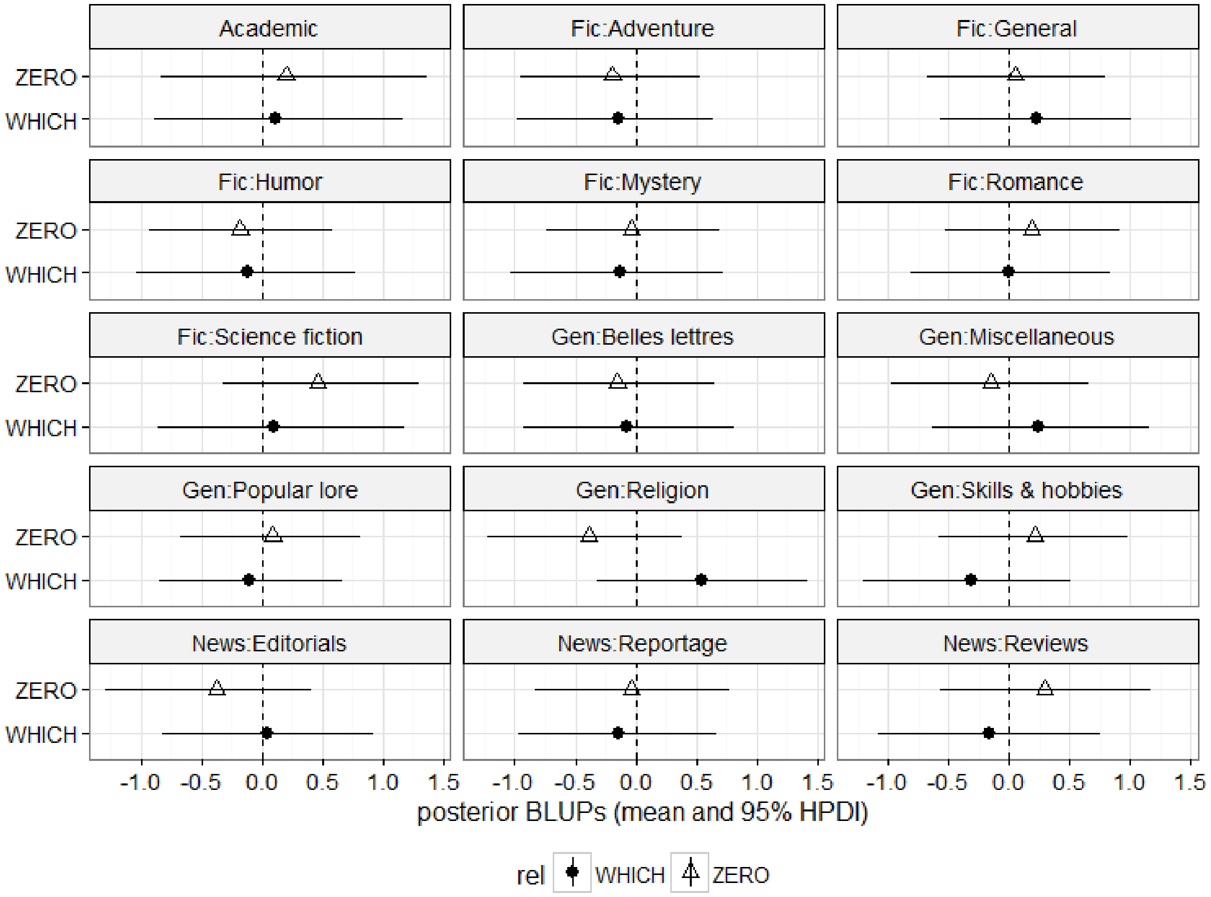


Figure A.5: Posterior means and 95% HPD intervals for adjustments to intercept by category in NSRC model (log odds scale). Estimates are compared to baseline log odds of *that*.

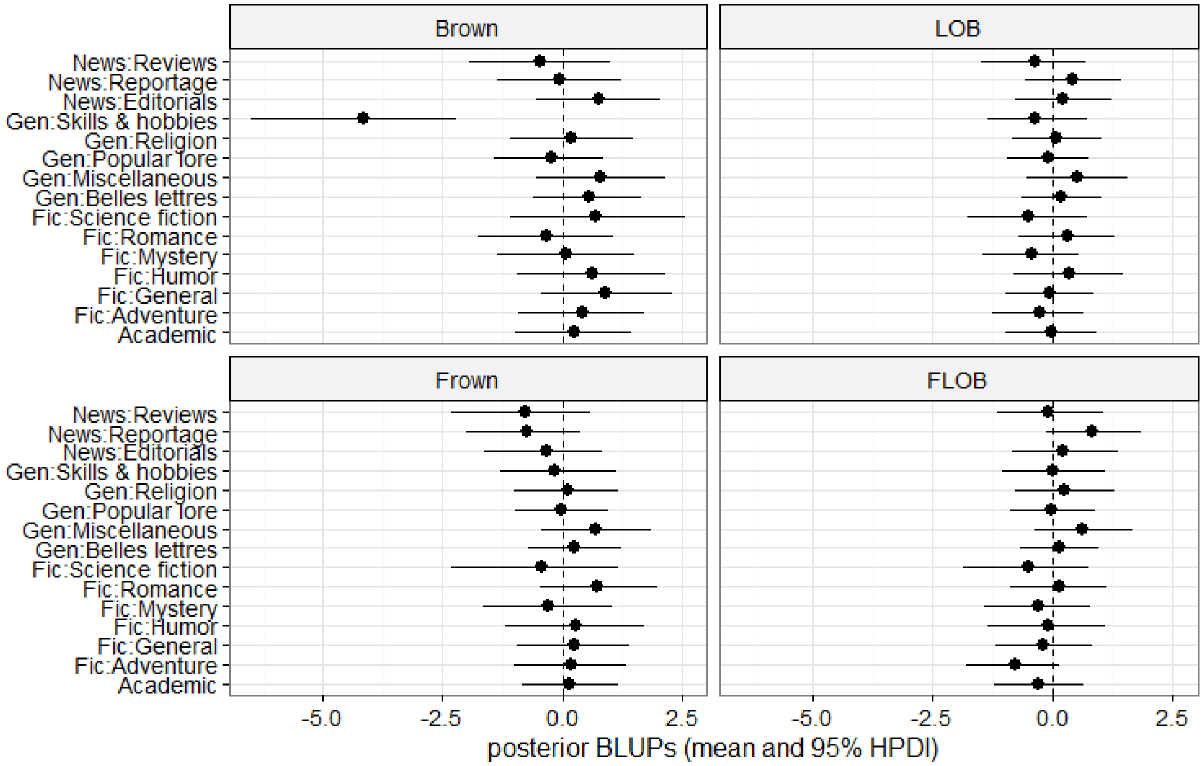


Figure 12: Posterior means and 95% HPDIs for adjustments to intercept by corpus and category in SRC model (log odds scale). Model predictions were for *which*.

References

Ball, Catherine N. 1996. A diachronic study of relative markers in spoken and written English. *Language Variation and Change* 8(2). 227–258. doi:10.1017/S0954394500001150.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Berlage, Eva. 2014. *Noun phrase complexity in English*. (Studies in English Language). Cambridge: Cambridge University Press.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2003. Compressed noun-phrase structures in newspaper discourse. In Jean Aitchison and Diana M. Lewis (eds.), *New Media Language*, 169–181. London: Routledge.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Clark, Eve V. and Tatiana V. Nikitina. 2009. One vs. more than one: Antecedents to plural marking in early language acquisition. *Linguistics* 47(1). 103–139. doi:10.1515/LING.2009.004.

D’Arcy, Alexandra and Sali A. Tagliamonte. 2010. Prestige, accommodation, and the legacy of relative *who*. *Language in Society* 39(03). 383–410. doi:10.1017/S0047404510000205.

D’Arcy, Alexandra and Sali A. Tagliamonte. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(03). 255–285. doi:10.1017/S0954394515000101.

Desmet, Timothy, Constantijn De Baecke, Denis Drieghe, Marc Brysbaert & Wietske Vonk. 2006. Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes* 21(4). 453–485. doi:10.1080/01690960400023485.

Fowler, H. W. and David Crystal. 2010. *A dictionary of modern English usage*. Pbk. ed. (Oxford World’s Classics). New York: Oxford University Press.

Fox, Barbara A. and Sandra A. Thompson. 2007. Relative Clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language* 31(2). 293–326. doi:10.1075/sl.31.2.03fox.

Francis, W. N. and H. Kucera. 1979. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, RI: Brown University.

Frazee, Joseph, Lars Hinrichs, Benedikt Szmrecsanyi and Axel Bohmann. 2015. Which-hunting and the Standard English relative clause: Online Supplement: Automatic Zero-Relative Detection. *Language* 91(4). s1–s3. doi:10.1353/lan.2015.0070.

Gennari, Silvia P. and Maryellen C. MacDonald. 2009. Linking production and comprehension processes: The case of relative clauses. *Cognition* 111(1). 1–23. doi:10.1016/j.cognition.2008.12.006.

Guy, Gregory R. and Robert Bayley. 1995. On the choice of relative pronouns in English. *American Speech* 70(2). 148–162. doi:10.2307/455813.

Hadfield, Jarrod D. 2010. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* 33(2). 1–22.

Hadfield, Jarrod D. 2015. MCMCglmm Course Notes. ms. https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

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. 437–474.

Hinrichs, Lars, Nicholas Smith and Bridget Waibel. 2010. Manual of information for the part-of-speech-tagged, post-edited “Brown” corpora. *ICAME Journal* 34. 189–231.

Hinrichs, Lars, Benedikt Szmrecsanyi and Axel Bohmann. 2015. *Which*-hunting and the Standard English relative clause. *Language* 91(4). 806–836.

Hundt, Marianne and Geoffrey Leech. 2012. Small is beautiful: On the value of standard reference corpora for observing recent grammatical change. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*, 175–188. Oxford: Oxford University Press.

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. 221–242.

Hundt, Marianne, David Denison and Gerold Schneider. 2012. Relative complexity in scientific discourse. *English Language and Linguistics* 16(02). 209–240. doi:10.1017/S1360674312000032.

Husson, Francois, Julie Josse, Sebastian Le and Jeremy Mazet. 2016. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.31.5.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62. doi:10.1016/j.cogpsych.2010.02.002 (3 May, 2015).

Jaeger, T. Florian. 2011. Corpus-based research on language production: Information density and reducible subject relatives. In Emily M. Bender and Jennifer E. Arnold (eds.), *Language from a Cognitive Perspective: Grammar, Usage, and Processing. Studies in honor of Tom Wasow*, 161–197. Stanford, CA: CSLI Publications.

Jaeger, T. Florian and Thomas Wasow. 2006. Processing as a source of accessibility effects on variation. In Rebecca T. Cover and Yuni Kim (eds.), *Proceedings of the 31st annual meeting of the Berkeley Linguistic Society*, 169–180. Ann Arbor, MI: Sheridan Books.

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

Keenan, Edward L. and Bernard Comrie. 1977. Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8(1). 63–99.

Labov, William. 1972. Some principles of linguistic methodology. *Language in Society* 1(01). 97. doi:10.1017/S0047404500006576.

Langacker, Ronald W. 2008. *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.

Leech, Geoffrey N., Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. (Studies in English Language). Cambridge, UK; New York: Cambridge University Press.

Lehmann, Hans Martin. 2001. Zero subject relative constructions in American and British English. *Language and Computers* 36(1). 163–177.

Levey, Stephen. 2006. Visiting London relatives. *English World-Wide* 27(1). 45–70. doi:10.1075/eww.27.1.04lev.

Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235-26. doi: 10.1515/cog-2015-0054.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In, *Advances in Neural Information Processing Systems*, 849–856. Cambridge, MA: MIT Press.

Lynch, Scott M. 2007. *Introduction to applied Bayesian statistics and estimation for social scientists*. (Statistics for Social and Behavioral Sciences). New York: Springer.

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4. 1–16. doi: 10.3389/fpsyg.2013.00226.

Mair, Christian. 2006. *Twentieth-century English: History, variation, and standardization*. (Studies in English Language). Cambridge, UK; New York: Cambridge University Press.

Mak, Willem M., Wietske Vonk and Herbert Schriefers. 2006. Animacy in processing relative clauses: The hikers that rocks crush. *Journal of Memory and Language* 54(4). 466–490. doi:10.1016/j.jml.2006.01.001.

Nevalainen, Terttu. 2012. Reconstructing syntactic continuity and change in Early Modern English regional dialects: The case of *who*. In David Denison, Ricardo Bermúdez-Otero, Chris McCully and Emma Moore (eds.), *Analyzing Older English*, 159–184. Cambridge: Cambridge University Press.

Pagès, Jérôme. 2014. *Multiple Factor Analysis by Example Using R*. (Chapman and Hall/CRC the R Series). Boca Raton, FL: CRC Press, Taylor and Francis Group.

Quirk, Randolph. 1957. Relative clauses in educated spoken English. *English Studies* 38. 97–109. doi:10.1080/00138385708596993.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. version 3.2.2. Vienna: R Foundation for Statistical Computing.

Rickford, John R. 2011. Relativizer omission in Anglophone Caribbean Creoles, Appalachian, and African American Vernacular English. In Emily Bender and Jennifer Arnold (eds.), *Language from a Cognitive Perspective: Grammar, Usage, and Processing. Studies in honor of Tom Wasow*, 139–160. Stanford, CA: CSLI Publications.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182. doi:10.1515/cogl.1996.7.2.149.

Rohdenburg, Günter. 2014. Relative clauses of reason in British and American English. *American Speech* 89(3). 288–311. doi:10.1215/00031283-2848978.

Roland, Douglas, Frederic Dick and Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57(3). 348–379. doi:10.1016/j.jml.2007.03.002.

Romaine, Suzanne. 1980. The relative clause marker in Scots English: Diffusion, complexity, and style as dimensions of syntactic change. *Language in Society* 9(02). 221–247. doi:10.1017/S004740450000806X.

Sand, Andrea and Rainer Siemund. 1992. 30 years on. *ICAME Journal* 16. 119–122.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. London; New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert and Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28(1). 1–29. doi:10.1017/S0954394515000198.

Tagliamonte, Sali. 2002. Variation and change in the British relative marker. In Patricia Poussa (ed.), *Relativisation on the North Sea Littoral*, 147–165. Munich: Lincom Europa.

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.

Temperley, David. 2003. Ambiguity avoidance in English relative clauses. *Language* 79(3). 464–484. doi:10.1353/lan.2003.0189.

Tottie, Gunnel and Dawn Harvie. 2000. It's all relative: Relativization strategies in early African American Vernacular English. In Shana Poplack (ed.), *The English history of African American English*, 198–230. Oxford: Blackwell.

Traxler, Matthew J., Rihana S. Williams, Shelley A. Blozis and Robin K. Morris. 2005. Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language* 53(2). 204–224. doi:10.1016/j.jml.2005.02.010.

Tweedie, Fiona J. and Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352. doi: 10.1023/A:1001749303137.

Wasow, Thomas, T. Florian Jaeger and David M. Orr. 2011. Lexical variation in relativizer frequency. In Horst J. Simon and Heike Wiese (eds.), *Proceedings of the 2005 DGfS workshop Expecting the unexpected: Exceptions in Grammar"*, 175–196. Berlin / New York: De Gruyter Mouton.

Wells, Justine B., Morten H. Christiansen, David S. Race, Daniel J. Acheson and Maryellen C. MacDonald. 2009. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology* 58(2). 250–271. doi:10.1016/j.cogpsych.2008.08.002.

Wiechmann, Daniel. 2015. *Understanding Relative Clauses: A Usage-based View on the Processing of Complex Constructions*. Berlin: de Gruyter Mouton.

Authors

Name: Jason Grafmiller

Affiliation: KU Leuven

Email: jason.grafmiller@kuleuven.be

Name: Benedikt Szmrecsanyi

Affiliation: KU Leuven

Email: benszm@kuleuven.be

Name : Lars Hinrichs

Affiliation : University of Texas at Austin

Email: larshinrichs@utexas.edu