Register in variationist linguistics

Benedikt Szmrecsanyi
KU Leuven

Benedikt Szmrecsanyi, Professor of Linguistics in the Quantitative Lexicology and Variational Linguistics research group at the Katholieke Universiteit (KU) Leuven, writes this article exploring the connections between register and variationist linguistics. He is involved with various large-scale research projects in areas such as probabilistic grammar, variationist sociolinguistic research, linguistic complexity, and dialectology/dialectometry. Szmrecsanyi's books include *Grammatical Variation in British English Dialects: A Study in Corpus-based Dialectometry* (2013, Cambridge) and *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech* (Szmrecsanyi & Wälchli 2014, Mouton de Gruyter). He is currently a principal investigator on a major grant-funded research project titled 'The register-specificity of probabilistic grammatical knowledge in English and Dutch', a project aimed at exploring the question of whether register differences lead to differences in the processes of making linguistic choices. In sharp contrast to the status quo in variationist linguistics, where register is often ignored entirely, much of Szmrecsanyi's variationist research treats register as a variable of primary importance. The findings from these studies have led Benedikt Szmrecsanyi to state that "we need more empirical/variationist work to explore the differences that register makes" (Szmrecsanyi 2017: 696).

Keywords: variationist sociolinguistics, corpus linguistics, probabilistic grammar

1. How is register conceptualized in variationist linguistics?

Variationist linguistics is a discipline in the field of variation studies that investigates variation between "alternate ways of saying 'the same' thing" (Labov 1972: 188). In this spirit, variationist linguists carefully account for competing variants and draw on quantitative methodologies to model the conditioning factors that regulate the way language users choose between semantically and functionally equivalent variants.

https://doi.org/10.1075/rs.18006.szm
Register Studies 1:1 (2019), pp. 76–99. issn 2542-9477 | e-issn 2542-9485

Variationist linguistics and traditional register research thus adopt different, not always compatible perspectives on language variation and its intersection with what we may call the register-genre-style triad. According to customary definitions (see, e.g., Biber & Conrad 2012: 8), analyzing registers means strictly speaking investigating the functional relationship(s) between a set of linguistic features and the situational context. Variationist linguistics, however, is about variation between different ways of accomplishing the same function (Labov 1972: 188). Therefore, functional variation does by definition not come under the remit of variationist linguistics because attention is typically restricted to functionally equivalent forms (see also Biber & Conrad 2004: 265 on this point). This is why variationist linguistics is largely agnostic about functional relationships.

Against this backdrop, it is maybe not surprising that there is quite a bit of terminological variance regarding the register-genre-style triad in the variationist literature: some variationists use the term 'register' (e.g. Gries 2015), some use the term 'genre' (e.g. Grafmiller 2014), and variationist sociolinguists in particular focus on 'style' (e.g. Eckert & Rickford 2001), which is often conceptualized as the amount of attention paid to one's speech. In the corpus-based variationist literature, the concept that most researchers really have in mind when they use the terms 'register' and 'genre' comes, a tad confusingly, close to what Biber and Conrad (2012: 2) refer to as 'style': aesthetic preferences that cause speakers and writers to favor particular variants in particular situations (and this also happens to be the conceptualization that underlies the case study to be presented later in this article). In actual research practice, of course, many corpus-based variationist linguists tend to rely on the register/genre distinctions built into the design of existing publicly available corpora (consider, e.g., the Brown corpus family with its 15-fold register/genre categorization – see Table 1 below) without losing much sleep over functional relationships. Sometimes, register is treated as an outright nuisance factor that needs to be controlled for in the analysis so as not to distort results. I am not aware of variationist work where the goal is the *discovery* of relevant register distinctions.

A second major difference in research philosophy between register research and variationist linguistics is that register analysts often aggregate over multiple features to characterize registers and register differences, the rationale being that it is hard to find individual features that comprehensively characterize registers (Biber & Conrad 2012: 9). By contrast, orthodox variationist analysis eschews aggregation and proceeds in what Nerbonne (2009: 176) calls a "single-feature-based" mode: researchers explore individual linguistic variables, one variable at a time (typically, one variable per research paper). It is certainly true that the field is moving towards multi-variable designs (see, e.g., Guy 2013, Guy & Hinskens 2016), but the one-variable-at-a-time approach is still the customary one.

A third point of difference is that while in traditional register research, the typical unit of observation is individual texts, variationist analysis is focused – as well shall see below – on individual linguistic choices. The task in variationist analysis is therefore to link register differences not to text frequencies but to linguistic choice-making.

In summary, then, in variationist linguistics register is typically conceptualized as stylistic variation in aesthetic preferences. Register is thought of as one of the language-external factors (beside, e.g., real time, geographic provenance, etc.) that regulates variation of individual linguistic variables. Register is specifically analyzed in terms of how it influences linguistic choices between functionally equivalent variants.

2. How does register relate to the research goals within variationist linguistics?

The central research goal in variationist linguistics is to understand how people choose between different ways of saying the same thing. Specifically, variationist linguists study the constraints that regulate linguistic choice-making. Constraints can be language-internal (e.g., the phonetic environment in choice contexts) or language-external (e.g., demographic factors, geography, etc.). Register, then, is one of the language-external factors that needs to be included in good models of how people choose between linguistic variants, for the sake of better understanding how language-internal and language-external constraints shape linguistic variation. I exemplify by considering the well-known alternation in the grammar of English between the prepositional dative construction, as in (1a), and the ditransitive dative construction, as in (1b):

(1) a. **I've sent a message to him** via a couple of different channels […]
(Corpus of Global Web-based English GB G)
b. **I sent him a message** saying I am simply going out with friends […]
(Corpus of Global Web-based English AU G)

Röthlisberger, Grafmiller, and Szmrecsanyi (2017) study this syntactic alternation based on corpus material sampling multiple geographic varieties of English and a range of registers. To understand how variation is constrained in the material at hand, the study fits a logistic regression model predicting dative choice based on well-known language-internal constraints (e.g., animacy of the recipient, length of the constituents) and two language-external factors, variety of English (e.g., British English versus Canadian English) and register (e.g., spoken-informal texts versus spoken-formal texts). Analysis shows that variety is a more important factor than register, but at the same the two factors interact significantly, which means that stylistic norms vary across varieties of English. Accordingly, in the Röthlisberger et al. (2017) study, register is quite central to the research goal of variationist linguistics: understanding how variation works.

With that being said, it is fair to say that register is a language-external factor that has received less attention in the variationist literature than other factors, such as social factors. Note in this connection, however, that the field is not entirely homogeneous: in Szmrecsanyi (2017), I draw a distinction between (Labovian) variationist sociolinguistics (also known as the 'Language Variation and Change paradigm'), and corpus-based variationist linguistics. Among other things, variationist sociolinguists are more interested in social determinants of variation than corpus-based variationist linguists, who typically focus more on language-internal constraints on variation. But also, because corpus-based variationist linguists typically rely on large, publicly available corpora that often sample multiple text types, register variation is a topic that is comparatively more important in corpus-based variationist linguistics. Sometimes register is conceptualized as a key factor of substantial interest (e.g. in Szmrecsanyi 2006, who investigates among other things how persistence/priming effects differ across registers), and sometimes register is modeled as a factor that may not be of primary interest but that needs nonetheless attention so as to not distort results and for the sake of accounting for variation (this, for instance, is the spirit of the Röthlisberger et al. 2017 study discussed above). But in either case, corpus-based variationist linguists do care about register variation. By contrast, it seems fair to say variationist sociolinguists of the Labovian persuasion tend to be less interested in register. It is true that style and style-shifting have received ample attention in variationist sociolinguistics (see e.g., Bell 1984, Rickford & McNair-Knox 1994). But, variationist sociolinguists rarely consider, e.g., non-spoken texts, and tend to be especially interested in vernacular speech as manifested in sociolinguistic interviews: the credo is, in a nutshell, that "variation in language is most readily observed in the vernacular of everyday life" (Tagliamonte 2012: 2; see D'Arcy & Tagliamonte 2015 for critical discussion). There are two reasons why vernacular speech has a special status in theorizing in variationist sociolinguistics: For one thing, vernacular speech is considered "the style in which the minimum attention is given to the monitoring of speech" (Labov 1972: 208), so the vernacular is where variation is thought to be at its best (unlike, as the reasoning goes, written language, which is more "governed by prescription"; D'Arcy & Tagliamonte 2015: 255). Second, more practically speaking, the bulk of work in variationist sociolinguistics deals with phonetic variables, where written texts and, to some extent, formal speech are irrelevant.

What have we learned about register and register variation through variationist linguistics? As far as the variationist sociolinguistic literature is concerned, there is plenty of evidence that style shifting across spoken styles, which can be seen as a form of register variation, is an excellent diagnostic of the social meaning of variation. As Rickford and Eckert (2001: 1) put it, "style is a pivotal construct in the study of sociolinguistic variation", which is why sociolinguistic interviews are designed to elicit a range of speech styles. Historically speaking, variationist sociolinguistic interest in style goes back to Labov (1966). As indicated above, Labov essentially defined style as the attention paid to one's speech. Style shifting, then, indicates the perceived prestige of variants. For example, in the famous department store study Labov (1966) found that many department store employees are more likely to pronounce postvocalic /r/ in careful, emphatic pronunciation, which demonstrates that rhoticity is prestigious. The generalization is that careful speech increases the rate of higher-prestige variants, while casual speech increases the rate of lower-prestige variants. At the same time, style-shifting interacts with age and sex: normally, there is more style-shifting in female speech than in male speech (Eckert 2000: 195). It should be added that more nuanced perspectives on style have been developed over the years, especially in what is known as third-wave sociolinguistics (Eckert 2018), but it is not clear that this work comes under the remit of variationist linguistics as defined here.

As to the corpus-based variationist literature, register has been shown time and again to be an important factor regulating variation. Recent studies that have found a significant main effect of register under multivariate control, or substantial random effects, include the following:

- Heylen (2005) investigates constituent order variation in varieties of German and finds that the difference between the spoken and written medium has a significant impact on variant choice;
- Grondelaers, Speelman, and Geeraerts (2008) model the Dutch postverbal *er* ('there') retention versus omission alternation and show that register (UseNet discourse versus popular newspapers versus quality newspapers) is a significant factor in Belgian Dutch;
- Levshina (2011: Chapter 6), in her study of the Dutch *doen* versus *laten* alternation, reports some significant register effects, with *doen* being particularly unlikely to be used in conversations and disfavored in web-based Dutch (Usenet discourse) in comparison to Dutch from newspapers;
- Lohmann (2011) analyzes the *help* versus *help to* alternation in English and finds that including genre distinctions significantly improves model accuracy;
- Wolk, Bresnan, Rosenbach, and Szmrecsanyi (2013) study the historical development of, among other things, the English dative alternation in A

Representative Corpus of Historical English Registers (ARCHER) and report that factoring in register differences improves model accuracy;

- Pijpops and Van De Velde (2014) find significant register effects (chat versus email versus quality newspaper versus tabloid) in some of their models of the Dutch partitive genitive alternation;
- Gries (2015) investigates the English particle placement alternation and finds that what he calls "sub-registers" (110) (e.g. private versus public dialogue, scripted versus unscripted monologue, and so on) matter for predicting particle placement choices;
- Rosemeyer and Enrique-Arias (2016) model the conditioning of variation in the expression of possession in Old Spanish and report that in their Bible corpus, register variation (lyrical, narrative) has a significant effect on variant choices;
- Szmrecsanyi et al. (2016) investigate ternary genitive variation in the late Modern English period and find that including register information (news versus science versus letters) as a random effect improves model accuracy;
- Heller (2017) is concerned with the genitive alternation across a range of postcolonial varieties of English and reports that the distinction between written and spoken registers has a significant effect on variant choice;
- Grafmiller and Szmrecsanyi (in press) study the particle placement alternation across varieties of English and report that according to conditional random forest analysis, register variation (written formal versus written informal versus spoken formal versus spoken informal) often has a substantial impact on variant choices.

The cumulative weight of evidence from research on style-shifting and corpus-based variationist linguistics thus suggests that register regularly[1] affects the relative frequency with which speakers and writers select particular variants. Specifically, register and style differences can be utilized as a diagnostic of prestige differentials, and in statistical models of grammatical variation based on corpus data, register distinctions often improve model quality.

[1] Failures to obtain significant register effects are not systematically reported in the literature, which is why it is hard to be more precise.

3. What are the major methodological approaches that are used to analyze or account for register in variationist linguistics?

To account for register in variationist linguistics, analysts apply the variationist method (see, e.g., Labov 1969, Sankoff 1988) and include register as (one of) the language-external constraints on the variation phenomenon under study. Cacoullos and Walker (2009: 326–327) concisely define the gist of the variationist method as follows:

> [the variationist method] seeks to discover patterns of usage in the relative frequency of co-occurrence of linguistic forms and elements of the linguistic context. The interpretative component of the variationist method lies in identifying similar discourse functions of different constructions […] We account for the selection of variants to fulfill a particular discourse function by exhaustively extracting each instance of that function in discourse and applying quantitative techniques to determine the influences of contextual factors on the choice of form.

In practice, conducting a variationist analysis of register effects consists of the following steps:

1. **Selection of the variable**: The analyst picks one (and traditionally, only one) variation phenomenon (also known as a linguistic variable or – in the realm of grammar – an alternation). If the analyst has a particular interest in how register shapes variation patterns, she will want to select a variation phenomenon that we have reason to believe is particularly sensitive to register differences and/or style-shifting.
2. **Circumscription of the variable context**: Guided by the Principle of Accountability ("any variable form […] should be reported with the proportion of cases in which the form did occur in the relevant environment, compared to the total number of cases in which it might have occurred", Labov 1969: 738, n. 20), the analyst catalogs all variant forms and properly circumscribes the variable context to ensure that attention is restricted to only those contexts in which language users truly have a choice between all competing variants.
3. **Retrieval and annotation**: Based on the circumscription of the variable context in step 2, the analysts next turns to production/corpus data and identifies and extracts all relevant variant forms in the material. Subsequently, the analyst annotates each variant form for language-internal and language-external constraints on variation. To identify the set of potentially relevant language-internal constraints, analysts typically survey the literature and/or rely on intuitions. The language-external constraints are typically determined on a

by-text or by-transcript basis (i.e., register, real time, demographic characteristics of the speaker/writer, and so on).

4. **Analysis:** The richly annotated dataset generated in step 3 is then analyzed statistically to determine the conditioning of variation using multivariate analysis methods such as logistic regression analysis or conditional random forest analysis (see, e.g., Tagliamonte & Baayen 2012 for an accessible introduction).
5. **Interpretation**: What do the results generated in step 4 reveal about the interaction between register and linguistic variation?

In variationist analysis, the unit of observation is thus individual linguistic choices, and not, e.g., texts (see Biber, Egbert, Gray, Oppliger, & Szmrecsanyi 2016 for discussion). Relevant research questions about register include the following: How do particular registers influence the odds that people select particular variants? How powerful a predictor is register vis-à-vis other language-external constraints? Also, in terms of explanatory power, how does register fare vis-à-vis language-internal constraints?

4. What does a typical register study look like in variationist linguistics?

4.1 Selection of the variable

As an empirical case study, we will now investigate variation between relativizer *which*, as in (2a), and relativizer *that*, as in (2b):

(2) a. The largest hurdle the Republicans would have to face is a **state law which says** that before making a first race, one of two alternative courses must be taken (Brown text A01)

b. Fulton legislators work with city officials to pass **enabling legislation that will permit** the establishment of a fair and equitable pension plan for city employees (Brown text A01)

Which and *that* in contexts such as (2) qualify as different ways of saying the same thing. The research question to be addressed in this case study is the following: How important a factor is register vis-à-vis other constraints in shaping variation between the explicit relativizers *which* and *that*? We know that *which-that* variation is subject to a number of language-internal probabilistic constraints. We also know that written English – particularly American English – is drifting towards increased usage of *that* because of a process that Hinrichs, Szmrecsanyi, and Bohmann (2015: 806) call “institutionally backed colloquialization”. Finally, we know that *which* is the incoming form that has a bookish feel to it, while *that* is the more colloquial variant that is widespread in vernacular language (Biber, Johansson, Leech, Conrad, & Finegan 1999: 610; Tagliamonte, Smith & Lawrence 2005). In summary, then, *which-that* variation is richly constrained by language-internal and language-external constraints, which makes this alternation an interesting phenomenon to study from a variationist perspective.

4.2 Circumscription of the variable context

I adopt the circumscription of the variable context used in Hinrichs et al. (2015). Attention is thus restricted to:

- finite relative clauses introduced by *which* and *that*, ignoring, e.g., participial relative clauses of the type *the man standing at the bar*;
- restrictive relative clauses, because standard English non-restrictive relative clauses (as in [3]) only allow *which*. The criterion for restrictiveness is the absence of a comma preceding the relativizer, which in standard written English is a sufficiently reliable diagnostic;
- subject relative clauses (i.e., clauses where the relativizer acts as subject), because restrictive non-subject relative clauses also permit zero as a relativizer (see [4]);
- relative clauses with inanimate antecedents, as animate antecedents tend to trigger the relativizers *who/whom/whose* (as in [5]).

Lastly, the study ignores oblique relatives with pied-piping (as in [6]) because these categorically take *which*.

(3) The jury further said in term-end presentments that the City Executive Committee, **which had over-all charge of the election,** deserves the praise and thanks of the City of Atlanta (Brown text A01)

(4) [...] that what we were asserting to be bad was precisely the suffering ∅ **we thought had occurred back there** [...] (Brown text J52)

(5) Barber, who is in his 13th year as a legislator, said there are **some members** of our congressional delegation in Washington **who would like to see it (the resolution) passed.** (Brown text A01)

(6) [...] and concentrate its constructive efforts on eliminating in other parts of Latin America the social conditions **on which totalitarian nationalism feeds** (Brown text A04)

4.3 Retrieval and annotation

We will re-analyze[2] a subset of the relativizer dataset analyzed in Grafmiller, Szmrecsanyi, and Hinrichs (2016), which, in turn, largely overlaps with the relativizer dataset analyzed in Hinrichs et al. (2015). The dataset is publicly available as supplementary materials to Grafmiller et al. (2016) <https://doi.org/10.1515/cllt-2016-0015>. Variable relativizer tokens were extracted from Brown, Frown, LOB, and F-LOB (see Hinrichs, Smith, & Waibel 2010 and references therein). Each corpus contains roughly 1 million words of written text of standard American (Brown, Frown) and British (LOB, F-LOB) English compiled in the early 1960s (Brown, LOB) and the early 1990s (Frown, F-LOB). Each corpus consists of 500 2,000-word samples representing data from 15 distinct 'categories' (in Brown parlance), e.g. newspaper articles, humor writing, academic writing, and various genres of fiction. These categories can be grouped into four 'genre groups' (again, in Brown parlance): (1) press, (2) general prose, (3) learned, and (4) fiction.

Extraction of variant forms was in line with the description of the variable context detailed above (in addition to subject relative clauses, the dataset also covers non-subject relative clauses, which however are not analyzed in the present study). To extract instances of relative clauses with overt relativizers, the compilers made extensive use of the Part-Of-Speech tagging available in the four corpora. For additional details on the extraction methods, see Hinrichs et al. (2015: 815–816). In total, the subject relativizer dataset to be analyzed in the present paper spans $N=4{,}400$ *which* tokens and $N=5{,}731$ *that* tokens.

All tokens were annotated for a suite of constraints thought to influence the choice of relativizer. In the present study, I consider a smaller set of constraints which, according to the analysis in Hinrichs et al. (2015) and Grafmiller et al. (2016), are particularly important:

- **Preceding relativizer (PRECREL).** Which relativizer was used last time the writer had a choice? The predictor distinguishes the levels 'that', 'which', 'zero', and 'none' (for when the relativizer in question is the first one encountered in a corpus file). Consider (7), where the choice context *the primes which enter* is preceded by the choice context *The theorem which we prove*.

> (7) **The theorem which we prove** is more general than what we have described since it works with the primary decomposition of the minimal polynomial whether or not **the primes which enter** are all of first degree.
> (Brown text J18)

[2] The analyses were carried out using the R software environment for statistical computing <https://www.r-project.org/>. The R script is available at <https://osf.io/k4rfp/>.

This predictor gauges the effect of priming or structural persistence (see, e.g., Gries 2005; Szmrecsanyi 2005): language users tend to recycle recently used grammatical variants in upcoming discourse. A univariate frequency analysis indicates that some 78.5% of *that* relativizers are preceded by another *that*, while 74.0% of *which* relativizers are preceded by another *which*.

- **Part of Speech of antecedent (ANTPOS).** The annotation distinguishes two levels, 'noun' vs. 'other', to gauge the effect of antecedent pronominality independently from definiteness, and to distinguish lexically specific antecedents from 'empty' ones, as in (8).

  (8) **all that happens** is that the better qualified teacher declines to gamble two or three years of his life on the chance that conditions at the Catholic institution will be as good as those elsewhere. (Brown text A35)

  The analysis in Hinrichs et al. (2015: Table 4) indicates that nominal antecedents favor *which*, and in the dataset subject to analysis here *which* shows larger occurrence rates when the antecedent is nominal (44.5%) than when it is non-nominal (35.3%). The reverse is true for *that* (55.5% versus 64.7%).

- **Length of antecedent in words (ANTLN).** This measure gauges the complexity of the noun phrase modified by the relative clause. Consider (9), where the antecedent (*the escheat law*) has a length of three words.

  (9) He told the committee the measure would merely provide means of enforcing **the escheat law which has been on the books "since Texas was a republic"**. (Brown text A02)

  The analysis in Hinrichs et al. (2015: Table 4) suggests that increasing antecedent length disfavors *that* and favors *which*. In the dataset under analysis here, antecedent of *that*-relative clauses have a mean length of 3.24 words, while *which*-antecedents have a mean length of 3.54 words.

- **Length of relative clause in words (RCLN).** This measure approximates the complexity of the clause introduced by the relativizer. Consider (9), where the relative clause (*which has been on the books "since Texas was a republic"*) has a length of 11 orthographic words. The analysis in Hinrichs et al. (2015: Table 4) suggests that increasing relative clause length disfavors *that* and favors *which*. In the dataset under analysis here, the mean length of *that*-relative clauses is 8.80 words, while *which*-relative clauses have a mean length of 10.22 words.

- **Passive-active ratio (PASSIVEACTIVERATIO).** This predictor measures the proportion of passive constructions (as in [10]) over active lexical verbs in a given corpus text.

  (10) His contention **was denied** by several bankers [...]. (Brown text A02)

  The analysis in Hinrichs et al. (2015: Table 4) suggests that increased usage of the passive voice correlates with reduced *that*-usage. Likewise, in the dataset under analysis here, the mean proportion of passive constructions is 0.40 for *that*-relatives, while it is 0.70 for *which*-relatives.

- **Time (TIME).** This is a binary predictor ('1960s' vs. '1990s') indicating the time period when the text was sampled. The analysis in Hinrichs et al. (2015: Table 4) shows that texts from the 1990s favor usage of *that*, compared to texts from the 1960s. In the dataset under analysis, the proportion of *that*-relatives to *which*-relatives was 46.7%: 53.3% in the 1960s, but 66.1%: 33.9% in the 1990s.

- **Variety (VARIETY).** This is a binary predictor indicating the variety of written Standard English ('American English' vs. 'British English') from which the text was sampled. The analysis in Hinrichs et al. (2015: Table 4) suggests that all other things being equal, American English strongly favors *that* compared to British English. In the dataset under analysis, the proportion of *that*-relatives to *which*-relatives is 71.5%: 28.5% in American English, but only 41.0%: 59.0% in British English.

- **Genre group (GENREGROUP).** This predictor distinguishes the following genre groups: (1) press, (2) general prose, (3) learned, and (4) fiction. In the dataset under analysis, the proportion of *that*-relatives to *which*-relatives is 72.5%: 27.5% in fiction, 54.5%: 45.5% in general prose, 43.4%: 56.6% in learned writing, and 61.1%: 38.9% in press writing.

- **Category (CATEGORY).** Relativizer proportions in the 15 categories covered in the Brown family are displayed in Table 1.

- **Corpus file (CORPUSFILEID).** This logs the particular corpus text in which a relativizer occurs, for the sake of modeling author idiosyncracies as a by-subject random effect in regression analysis.

Table 1. Variant rates by Brown category

	that	*which*	Total
A_Reportage	393	286	679
	57.9%	42.1%	6.7%
B_Editorial	304	214	518
	58.7%	41.3%	5.1%
C_Review	264	113	377
	70.0%	30.0%	3.7%
D_Religion	272	268	540
	50.4%	49.6%	5.3%
E_Skills	500	244	744
	67.2%	32.8%	7.3%
F_Popularlore	624	370	994
	62.8%	37.2%	9.8%
G_BellesLettres	901	829	1730
	52.1%	47.9%	17.1%
H_Miscellaneous	226	399	625
	36.2%	63.8%	6.2%
J_Science	895	1165	2060
	43.4%	56.6%	20.3%
K_GeneralFiction	286	127	413
	69.2%	30.8%	4.1%
L_Mystery	253	78	331
	76.4%	23.6%	3.3%
M_ScienceFiction	77	29	106
	72.6%	27.4%	1.0%
N_Adventure	383	116	499
	76.8%	23.2%	4.9%
P_Romance	251	98	349
	71.9%	28.1%	3.4%
R_Humor	102	64	166
	61.4%	38.6%	1.6%
Column Total	5731	4400	10131

4.4 Analysis

The distributional analyses provided in the foregoing discussion suggest that we are dealing with substantial variation between relativizer *which* and *that*, conditioned by both language internal-and language-external probabilistic constraints. The task before us is to model this variation statistically to determine the overall importance, effect direction, and effect size of constraints on relativizer variation under multivariate control. Multivariate control is essential to make sure that, e.g., the high rate of *which* (63.8%) in the Miscellaneous category (see Table 1) is not trivially due to the fact that relative clauses also happen to be longest (10.8 words on average, versus a global mean of 9.4 words) in the Miscellaneous category. Multivariate analysis techniques can take care of such confounds.

To evaluate the overall importance of the constraints, I begin by submitting the dataset to conditional random forest (CRF) analysis as implemented in the cforest() function in R's party package (Hothorn, Hornik & Zeileis 2006; Strobl, Boulesteix, Kneib, Augustin, & Zeilis 2008; Strobl, Boulesteix, Zeileis, & Hothorn 2007). I skip a discussion of the technicalities and instead refer the reader to the very accessible introduction in Tagliamonte and Baayen (2012). What is especially appealing from the point of view of the present study is that CRF can be used to straightforwardly rank constraints according to their explanatory importance. This is the purpose of the dot plot in Figure 1, which ranks constraints on relativizer variation according to their importance.

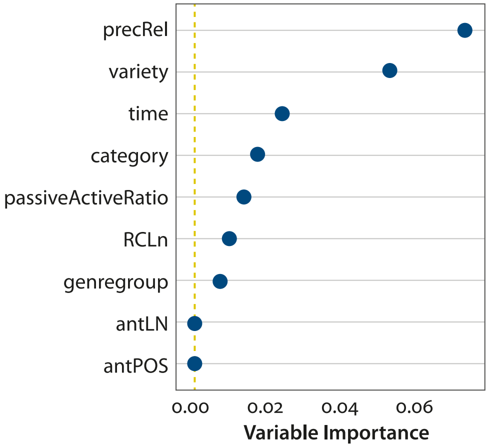

Figure 1. Conditional random forest permutation variable importance measures for constraints on *that* versus *which* variation. $C = 0.90$

According to CRF analysis, the most important constraint on relativizer variation is PRECREL, the language-internal predictor that checks which relativizer was used last time the writer had a choice. Strong persistence/priming effects in grammatical variation have been reported before in the literature (e.g., Scherre & Naro 1991; Weiner & Labov 1983), but it is noteworthy that the nature of the preceding relativizer (PRECREL) is the single most important predictor of relativizer choice. The next three predictors in the ranking are all language-external in nature: VARIETY (British English versus American English), TIME (1960s versus 1990s), and CATEGORY (which distinguishes between the 15 registers displayed in Table 1). In other words, regional differences are the most important language-external constraint, while register differences are least important – though it should be stressed that registers differences are still more central than most of the language-internal predictors in the model except PRECREL. For exploratory reasons, I also included the rather coarse-grained predictor GENREGROUP (press versus general prose versus learned versus fiction) in the analysis, but Figure 1 shows that its importance is limited. This appears to suggest that we really do need the full 15-fold distinction in the predictor CATEGORY to responsibly model register variation. The predictor PASSIVEACTIVERATIO scores a bit higher than relative clause length (RCLN). Lastly, according to CRF analysis properties of the antecedent (ANTPOS and ANTLN) are fairly insubstantial.

Having thus determined the overall importance of the constraints under study, I now turn to mixed-effects binary logistic regression analysis (Pinheiro & Bates 2000) as implemented in R's lme4 package (Bates, Mächler, Bolker, & Walker 2015) to learn more about effect directions and effect strengths. Logistic regression analysis has been the workhorse analysis technique in variationist linguistics for decades – the Varbrul program (Cedergren & Sankoff 1974) is essentially an implementation of logistic regression. Logistic regression modeling is the closest a corpus analyst can come to conducting a controlled experiment; the technique models the combined contribution of all the factors considered in the analysis, systematically testing the probabilistic effect of each factor while holding the other factors in the model constant. In addition to so-called 'fixed effects', which are classically estimated predictors that assess the reliability of the effect of repeatable characteristics, mixed-effects modeling also includes 'random effects' to capture variation dependent on open-ended, potentially hierarchical and unbalanced groups (see Wolk et al. 2013: 399–400 for more discussion). As fixed effects, I entered all of the language-internal factors described above plus VARIETY and TIME; as random effects, I included intercept adjustments for CATEGORY (to gauge register differences) and CORPUSFILEID (to approximate author idiosyncrasies as a so-called by-subject effects). Note here that the analyst has some leeway in deciding which factor(s) should be integrated into the fixed-effects structure, and which into the random-effects structure. In the case study at hand, it would certainly have been possible to model, e.g., VARIETY as a random effect (after all, the British/American distinction is not necessarily repeatable, as the next study down the road may include, e.g., Australian English), or possibly CATEGORY as a fixed effect. But beside conceptual considerations (is a predictor repeatable or not?), more pragmatic considerations often play a role; regardless of the conceptual status of the predictor in question, the analyst may be justified in modeling a predictor as a random effect, in order not to waste precious degrees of freedom (see, e.g., Zuur, Ieno, Walker, Saveliev, & Smith 2009: 106 for discussion). Because the predictor

CATEGORY has no fewer than 15 levels, modeling the predictor as a random effect is an elegant solution. The downside is that random effect modeling cannot determine if differences between levels are statistically significant or not.

Coefficients associated with the fixed effects in the resulting model are displayed in Table 2. (The models omits the predictor GENREGROUP, which CRF analysis has shown to be rather dispensable).

Table 2. Coefficients of fixed effects in mixed-effects binary logistic regression analysis of variation between *that* versus *which*

	Coefficient (b)	
(Intercept)	−2.61	***
precRel (default: none)		
that	−0.32	***
which	0.37	***
zero	−0.01	
antPOS (default: non-nominal)		
Nominal	0.59	***
antLN	0.06	***
RCLn	0.05	***
passiveActiveRatio	0.50	***
variety (default: American English)		
British English	2.04	***
time (default: 1961)		
1991	−1.3	***

Note. Significance codes: 0 – '***'; 0.001 – '**'; 0.01 – '*'. Predicted odds are for *which*. *C* = 0.93; condition number (kappa) = 7.8;% outcomes correctly predicted: 80.6% (baseline: 56.6%).

The column 'Coefficient (b)' displays regression coefficients. Negative coefficients disfavor the predicted outcome, which is usage of *which*; positive coefficients favor the predicted outcome. Let us go through the coefficients one by one. The reference level for PRECREL is 'none' (this happens when the relativizer under analysis is the first in a given corpus text). Compared to this condition, a preceding *that* relativizer significantly disfavors usage of *which*; but a preceding *which* relativizer significantly favors re-use of *which* (a preceding zero relativizer does not have a significant effect on choice between *which* and *that*). Needless to say, these effects are perfectly in line with the literature on priming. Likewise consistent with the literature (Hinrichs, Szmrecsanyi & Bohmann 2015), nominal antecedents (ANTPOS), long antecedents (ANTLN), and long relative clauses (RCLN)

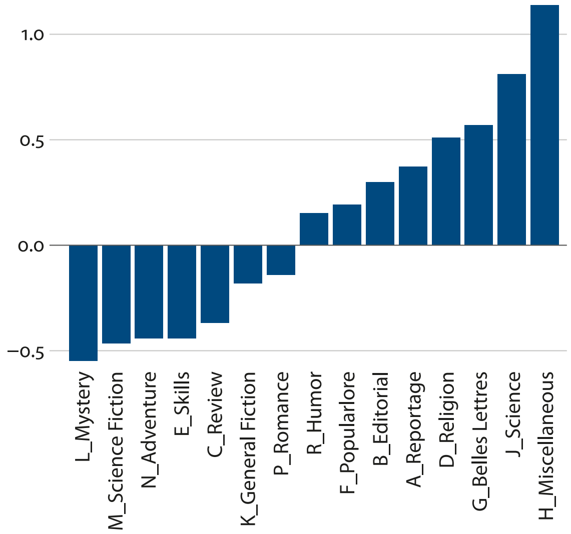

Figure 2. Intercept adjustments for random effect CATEGORY. Positive adjustments favor *which*, negative adjustments favor *that*

all favor choice of *which*. Also, increased usage of the passive voice (PASSIVE-ACTIVERATIO) increases the odds for *which* (in line with Hinrichs, Szmrecsanyi, & Bohmann 2015). As to the language-external predictors, observe that in comparison to American English, British English favors *which*, while texts written in the 1990s disfavor *which*, compared to texts written in the 1960s. In other words, we see a real time drift towards *that*, as described by Hinrichs, Szmrecsanyi & Bohmann (2015).

Let us finally address the issue most germane to this article: What is the effect that register distinctions have on relativizer choice? Figure 2 displays the intercept adjustments (i.e., adjustments to the base probability that the relativizer *which* is chosen, all other things being equal) that the model makes for individual levels of the random effect CATEGORY. The upshot is that the register categories that most strongly favor *which* are – in descending order of magnitude – Miscellaneous (H), Science (J), and Belles Lettres (G). The register categories that disfavor *which* and instead favor *that* are Mystery and detective Fiction (L), Science Fiction (M), and Adventure and Western (N). To come back to an issue mentioned earlier, then, Figure 2 provides multivariate evidence that, e.g., the high rate of *which* in the Miscellaneous category (see Table 1) is indeed *not* trivially due to the fact that relative clauses also happen to be longest in Miscellaneous: There really is a stylistic preference for *which* in that category.

4.5 Interpretation

As to language-internal constraints on *which-that* variation, the most important factor is structural persistence: There is a surprisingly strong tendency to re-use the relativizer which was used last time there was a choice, in line with the literature on priming and related phenomena (see Szmrecsanyi 2006 for extended discussion). The effect of most of the other language-internal constraints under consideration in the present study are consistent with Rohdenburg's complexity principle (Rohdenburg 1996), which stipulates that more explicit grammatical variants are used in cognitively more complex environments. Note now that *which* can be argued to be somewhat more explicit than *that*, thanks, among other things to, the fact that it contains more phonetic material than *that* ([wɪtʃ] or even [hwɪtʃ] versus [ðət]). And as we have seen in regression analysis, it is *which* that is favored when the antecedent is long (and thus arguably complex) and when the relative clause is long (and thus arguably complex). The positive correlation between the predictor PASSIVEACTIVERATIO and usage of *which* suggests that the increasing popularity of *that* is not primarily an outcome of the fact that 20th century style guides prescribe *that* in restrictive relative clauses ("Careful writers […] go which-hunting, remove the defining *whiches* and by so doing improve their work"; Strunk & White 1999: 59). Instead, the variation patterns seem to be regulated by formality (see Hinrichs, Szmrecsanyi, & Bohmann 2015 for discussion): *which*-users tend to consistently opt for formal linguistic options.

The hierarchy of importance (Figure 1) of the three language-external constraints we studied is variety > time > register. Register differences are thus less important than geographic contrasts or real time drifts, but on the other hand register differences are more important than most of the language-internal constraints except for PRECREL (priming). Analysis of the effect directions shows that the real-time drift towards relative *that* is significant, all other things being equal, and that it is led by American English.

With regard to register differences, we see that writers have aesthetic preferences such that informational registers (e.g., science texts) are particularly hospitable towards *which*, while imaginative registers (e.g., mystery and detective fiction) are particularly hospitable towards *that*. This distribution certainly supports the established view in the literature (e.g., Biber et al. 1999: 610; Tagliamonte, Smith, & Lawrence 2005) that *that* is the more vernacular variant, compared to 'bookish' *which*. Now, as we saw in the introductory section, interpreting variationist findings in terms of functional relationships is problematic because the variationist methodology *stricto sensu* requires analysts to restrict attention to functionally equivalent constructions and forms. But that being said, one could reasonably argue that the function of *that* is to establish a less formal, more vernacular tone, while the function of *which* is to establish a more formal, 'bookish' tone.

5. What are the most promising areas of future register research in variationist linguistics?

The research record suggests that register/genre/style distinctions are important ingredients in variationist accounts of the relative frequencies with which speakers and writers use linguistic variants. In the case study reported in the previous section, register happens to be a slightly less important factor overall than other language-external predictors (real time and regional differences), but other alternations may be differentially sensitive to register distinctions – further research into this issue is clearly needed.

Also, variationist linguists are going to have to address deeper theoretical questions about the register-genre-style triad. It is possible, and indeed likely, that the way that register has been modeled in the bulk of the previous literature is simplistic, in that register is typically modeled merely as a main effect (as was done in the case study in this article). This implicitly assumes that, yes, particular registers may favor usage of particular variants, but also that the way language-internal constraints (e.g., length effects) regulate variation stays constant across registers. But, we do not know about the extent to which language users, when they have the choice between different ways of saying the same thing, draw on different choice-making processes in different registers.

This issue is loaded theoretically but under-investigated empirically. In the variationist sociolinguistic community, there is a sense that variation grammars are stable across styles:

> For the most part, stylistic variation is quantitatively simple, involving raising or lowering the selection frequency of socially sensitive variables without altering other grammatical constraints on variant selection; indeed, it is commonly assumed in VR [Variable Rule, BS] analyses that the grammar is unchanged in stylistic variation. (Guy 2005: 562; see also Labov 2010: 265)

Guy's claim may indeed be correct if attention is restricted to style-shifting in sociolinguistic interviews. It is not clear, however that this quantitative simplicity easily generalizes to 'real' register variation, e.g., variation across speech and writing (see Szmrecsanyi 2017 for discussion), and so, it is odd that corpus-based variationist linguists have tended to ignore this issue. A notable exception to this negligence is Grafmiller (2014), who, based on corpora such as the Switchboard corpus and the Boston University Noun Phrase Corpus, investigates the English genitive alternation, as in (1d1).

(11) a. I have in this lecture sketched alternatives to the standard way of narrating **philosophy 's history** as part of the development of Western culture […] (COCA 2017 ACAD)

b. He seems to have a partial understanding of its place in **the history of philosophy** […] (COCA 2017 ACAD)

The genitive alternation is conditioned by a range of semantic, syntactic and phonological constraints. Grafmiller models these constraints across some six different registers/genres (conversation, learned writing, non-fiction, general fiction, western fiction, press). His analysis uncovers significant interactions between register and the probabilistic weights that language-internal constraints on genitive variation have. In model 1 (p. 482), for example, no less than five of the nine language-internal constraints under study (possessor animacy, possessor givenness, possessor/possessum length, type-token ratio, possessor text frequency) turn out to have significantly different effect sizes across registers. This can be interpreted as evidence that grammar is *not* unchanged in register variation and that language users do indeed adjust their choice-making processes to the situational context.

If we reasonably assume that English genitive variation is not entirely atypical, massive register specificity à la Grafmiller (2014) raises important theoretical and methodological questions about the nature and scope of grammatical variation, and about the interaction of this variation with socioculture. Among other things, if it is indeed the case that different registers come with differently sized constraints, then Guy's Grammatical Difference Hypothesis (Guy 2015), according to which having different (or differently sized) constraints means having a different grammars, would arguably warrant us to conclude that language users have a number of different register-specific variable grammars, akin to situations of diglossia or bilingualism.

Work is underway in the variationist community to investigate this issue more fully. For example, researchers at the KU Leuven and at the University of Birmingham (Principal Investigators: Benedikt Szmrecsanyi, Jason Grafmiller, Freek Van de Velde) have embarked in 2018 on a project entitled "The register-specificity of probabilistic grammatical knowledge in English and Dutch" (funded by the Research Foundation Flanders under grant # G0D4618N). The project investigates parallel grammatical alternations in English and Dutch. Based on corpus study and on rating task experiments, the project seeks to establish the extent to which language users adjust their choice-making to the situational context: What are the relevant register-related dimensions of variability: formal versus informal (formality), or written versus spoken (medium)? Do languages such as English and Dutch differ in terms of the importance of probabilistic register differences? And, what are the probabilistic constraints that tend to have particularly variable probabilistic effects across registers?

These are the type of questions where the theoretically responsible intersection of register studies and variationist linguistics can substantially advance theorizing in usage-based linguistics.

Acknowledgements

Thanks go to Alexandra Engel, Laura Rosseel, and Freek Van de Velde for helpful comments on an earlier draft of this paper.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145. https://doi.org/10.1017/S004740450001037X

Biber, D., & Conrad, S. (2004). Corpus-based comparisons of registers. In C. Coffin, A. Hewings, & K. O'Halloran (eds.), *Applying English grammar: Functional and corpus approaches* (pp. 40–56). London: Hodder Arnold.

Biber, D., & Conrad, S. (2012). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., Gray, B., Oppliger, R., & Szmrecsanyi, B. (2016). Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In M. Kytö & P. Pahta (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 351–375). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139600231.022

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Cacoullos, R., & Walker, J. A. (2009). The present of the English future: Grammatical variation and collocations in discourse. *Language*, *85*(2), 321–354. https://doi.org/10.1353/lan.0.0110

Cedergren, H., & Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, *50*(2), 333. https://doi.org/10.2307/412441

D'Arcy, A., & Tagliamonte, S. A. (2015). Not always variable: Probing the vernacular grammar. *Language Variation and Change*, *27*(3), 255–285. https://doi.org/10.1017/S0954394515000101

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High* (Language in Society 27). Malden, MA: Blackwell.

Eckert, P. (2018). *Meaning and linguistic variation: The third wave in sociolinguistics*. Cambridge: Cambridge University Press.

Eckert, P., & Rickford, J. R. (2001). *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.

Grafmiller, J. (2014). Variation in English genitives across modality and genres. *English Language and Linguistics*, *18*(3), 471–496. https://doi.org/10.1017/S1360674314000136

Grafmiller, J., & Szmrecsanyi, B. (in press). Mapping out particle placement in Englishes around the world. A case study in comparative sociolinguistic analysis. *Language Variation and Change*.

Grafmiller, J., Szmrecsanyi, B., & Hinrichs, L. (2016). Restricting the restrictive relativizer. *Corpus Linguistics and Linguistic Theory*, *14*(2), 309–355 https://doi.org/10.1515/cllt-2016-0015. <https://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2016-0015/cllt-2016-0015.xml> (1 March, 2018).

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399. https://doi.org/10.1007/s10936-005-6139-3

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, *10*(1), 95–125. https://doi.org/10.3366/cor.2015.0068

Grondelaers, S., Speelman, D., & Geeraerts, D. (2008). National variation in the use of er "there". Regional and diachronic constraints on cognitive explanations. In G. Kristiansen, & R. Dirven, *Cognitive sociolinguistics: Language variation, cultural models, social systems* (pp. 153–204). Berlin: De Gruyter. https://doi.org/10.1515/9783110199154.2.153

Guy, G. R. (2005). Letters to language. *Language*, *81*(3), 561–563. https://doi.org/10.1353/lan.2005.0132

Guy, G. R. (2013). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics*, *52*, 63–71. https://doi.org/10.1016/j.pragma.2012.12.019

Guy, G. R. (2015). *Coherence, constraints and quantities*. Talk given at NWAV44, Toronto.

Guy, G. R., & Hinskens, F. (2016). Linguistic coherence: Systems, repertoires and speech communities. *Lingua*, 172–173, 1–9. https://doi.org/10.1016/j.lingua.2016.01.001

Heller, B. (2017). Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. Unpublished PhD dissertation, KU Leuven. https://doi.org/10.1177/0075424216685405

Heylen, K. (2005). Zur Abfolge (pro)nominaler Satzglieder im Deutschen. Eine korpusbasierte Analyse der relativen Abfolge von nominalem Subjekt und pronominalem Objekt im Mittelfeld. Unpublished PhD dissertation, KU Leuven.

Hinrichs, L., Smith, N., & Waibel, B. (2010). Manual of information for the part-of-speech tagged, post-edited "Brown" corpora. *ICAME Journal*, *34*, 189–231.

Hinrichs, L., Szmrecsanyi, B., & Bohmann, A. (2015). Which-hunting and the Standard English relative clause. *Language*, *91*(4), 806–836. https://doi.org/10.1353/lan.2015.0062

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, *45*, 715–762. https://doi.org/10.2307/412333

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia, PA: University of Philadelphia Press.

Labov, W. (2010). *Principles of linguistic change. Vol. 3: Cognitive and cultural factors* (Language in Society 39). Malden, MA: Wiley-Blackwell. https://doi.org/10.1002/9781444327496

Levshina, N. (2011). Doe wat je niet laten kan [Do what you cannot let]: A usage-based analysis of Dutch causative constructions. Unpublished PhD dissertation, KU Leuven.

Lohmann, A. (2011). Help vs help to: A multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics*, *15*(3), 499–521. https://doi.org/10.1017/S1360674311000141

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, *3*(1), 175–198. https://doi.org/10.1111/j.1749-818X.2008.00114.x

Pijpops, D., & Van de Velde, F. (2014). A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory*, *10*, 1–30. https://doi.org/10.1515/cllt-2013-0027. <https://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2013-0027/cllt-2013-0027.xml> (14 February, 2018).

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-{PLUS}*. New York: Springer. https://doi.org/10.1007/978-1-4419-0318-1

Rickford, J. R., & Eckert, P. (2001). Introduction: John R. Rickford and Penelope Eckert. In P. Eckert & J. R. Rickford (Eds.), *Style and Sociolinguistic Variation* (pp. 1–18). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511613258.001. <http://ebooks.cambridge.org/ref/id/CBO9780511613258A010> (31 December, 2017).

Rickford, J. R., & McNair-Knox, F. (1994). Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. In D. Biber & E. Finegan (Eds.), *Perspectives on register: Situating register variation within sociolinguistics* (pp. 235–276). Oxford: Oxford University Press.

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

Rosemeyer, M., & Enrique-Arias, A. (2016). A match made in heaven: Using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish. *Language Variation and Change*, *28*(3), 307–334. https://doi.org/10.1017/S0954394516000120

Röthlisberger, M., Grafmiller, J., & Szmrecsanyi, B. (2017). Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics*, *28*(4), 673–710. https://doi.org/10.1515/cog-2016-0051

Sankoff, D. (1988). Sociolinguistics and syntactic variation. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge survey* (pp. 140–161). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620577.009

Scherre, M., & Naro, A. (1991). Marking in discourse: "Birds of a feather." *Language Variation and Change*, *3*, 23–32. https://doi.org/10.1017/S0954394500000430

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Strunk, W., & White, E. B. (1999). *The elements of style*, 4th ed. Longman.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113–150. https://doi.org/10.1515/cllt.2005.1.1.113

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197808

Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511763380

Szmrecsanyi, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *62*(4), 1–17. https://doi.org/10.1017/cnj.2017.34

Szmrecsanyi, B., Biber, D., Egbert, J., & Franco, K. (2016). Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change*, *28*(1), 1–29. https://doi.org/10.1017/S0954394515000198

Szmrecsanyi, B., & Wälchli, B. (Eds.). (2014). *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110317558

Tagliamonte, S. (2012). *Variationist sociolinguistics change, observation, interpretation*. Malden, MA: Wiley-Blackwell. <http://public.eblib.com/EBLPublic/PublicView.do?ptiID=819316> (29 August, 2013).

Tagliamonte, S., Smith, J., & Lawrence, H. (2005). No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change*, *17*(1), 75–112. https://doi.org/10.1017/S0954394505050040

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135–178. https://doi.org/10.1017/S0954394512000129

Weiner, J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*, 29–58. https://doi.org/10.1017/S0022226700007441

Wolk, C., Bresnan, J., Rosenbach, A., & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, *30*(3), 382–419. https://doi.org/10.1075/dia.30.3.04wol

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer. https://doi.org/10.1007/978-0-387-87458-6

Address for correspondence

Benedikt Szmrecsanyi
Department of Linguistics
KU Leuven
Blijde-Inkomststraat 21
B-3000 Leuven
Belgium
benszm@kuleuven.be