English Language and Linguistics, 28.1: 95–133. © The Author(s), 2023. Published by Cambridge University Press
doi:10.1017/S1360674323000497

Expanding the scope of grammatical variation: towards a comprehensive account of genitive variation across registers

DOUGLAS BIBER
Northern Arizona University
BENEDIKT SZMRECSANYI
KU Leuven
RANDI REPPEN and TOVE LARSSON
Northern Arizona University
(Received 19 April 2023; revised 20 August 2023)

Most studies of genitive variation in English have considered only the choice of two variants (*'s* versus *of*), based on analysis of only tokens that are judged to be interchangeable. We argue in the present article that research on genitive variation can be usefully extended in both respects: including premodifying nouns as a third variant; and attempting to account for *all* tokens of the genitive. In addition, we extend the scope of analysis to explore the possibility of contextual constraints having different importance in different registers.

First, we carry out a text-linguistic analysis comparing the rates of genitive variants in texts from three registers (conversation, newspaper reports, academic articles), showing that genitives overall are much more frequent in written registers, with the premodifying noun variant being especially frequent. Then, a variationist analysis is undertaken to account for the choice of genitive variant in particular contexts and registers. A total of 3,425 genitive tokens were coded for ten contextual characteristics (e.g. length of the Modifying NP, semantic category of the Modifying noun and the Head noun, final sibilancy of the Modifying noun). Statistical analyses with random forests and conditional inference trees are triangulated, showing how contextual factors interact in predicting the use of each genitive variant – and how patterns of variation differ across registers.

Keywords: genitives, premodifying nouns, text-linguistic versus variationist analyses, register variation

1 Introduction

The principle of accountability is arguably the foundation of variationist (socio)linguistics (see, e.g., Labov 1969: 737–8, fn. 20; 1972: 72), including both the sociolinguistic Language Variation and Change framework (LVC; see, e.g., Tagliamonte 2011) and the Corpus-based Variationist Linguistics framework (CVL; see Szmrecsanyi 2017). The principle of accountability mandates that a full analysis of linguistic variation will account for all tokens of all related linguistic variants:

> Accountability requires that all the relevant forms in the subsystem of grammar that you have targeted for investigation, not simply the variant of interest, are included in the analysis. (Tagliamonte 2011: 10)

Adherence to the principle of accountability is one of the main characteristics that distinguishes variationist research from much other research in corpus linguistics:

> Corpus linguistics typically uses counts per X number of words, e.g. 10 per thousand, 10 per 10,000 etc. However, in LVC research it is critical to know how a variant is influenced by a particular type of context compared to another. This requires knowing the distribution of a feature (variant) out of the total number of contexts where it could have occurred but did not. (Tagliamonte 2011: 19)

> CVL research observes the Principle of Accountability (Labov 1969: 738), hence it is linguistic choice-making processes and not text frequencies that take centre stage. (Szmrecsanyi 2017: 3)

That is, variationist research is committed to comparative analysis of each related linguistic variant (e.g. the choice between *will* and *BE going to*), while corpus researchers usually analyze the rates of occurrence for linguistic features (e.g. modal verbs).

The principle of accountability is usually discussed in conjunction with the methods that are deemed to be appropriate for quantitative analyses of linguistic variation, employing the construct of the *linguistic variable*. In its simplest conceptualization, a linguistic variable is operationally defined to capture 'alternate ways of saying "the same" thing' (Labov 1972: 188; see also Tagliamonte 2011: 4; Szmrecsanyi 2017: 3). Thus, the principle of accountability is usually discussed in relation to the definition of a linguistic variable:

> variable rules … depend upon a much more general and important principle of accountability which is required in the analysis of linguistic behavior: *that any variable form (a member of a set of alternative ways of 'saying the same thing') 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: 737–8, fn. 20; emphasis added)

> According to the principle of accountability, it is necessary to circumscribe the data to only those contexts that are functionally parallel as well as variable. (Tagliamonte 2011: 10)

However, early variationist publications were linguistically accountable in a more comprehensive sense: they accounted for the distribution of *all* variants in *all* contexts. The initial motivation for the development of the *linguistic variable* was to quantitatively account for the 'free' or 'optional' variation that was left over after the application of more traditional linguistic analyses. That is, traditional methods of analysis were employed to account for the patterns of linguistic variation that could be fully described in categorical terms, and statistical analyses were then undertaken to account for the remaining patterns of variation. Labov's (1969) analysis of contraction/deletion of the English copula is a clear example of this type, where section 3 of the

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

article provides a detailed description of the linguistic contexts that categorically prohibit contraction/deletion of the copula (e.g. in clause-final position). The subsequent variable rule analysis was designed to account for the remaining variation: contexts where the copula could be optionally contracted/deleted. The underlying motivation for the development of variable rules was to essentially argue that so-called 'free' or 'optional' variation was in fact predictable, at least in probabilistic terms. Thus, Labov criticized the conventions of traditional phonological rules, which used the () notation to mark optional elements, as 'no more useful … than the label "free variation"' (1969: 737). And he explicitly characterized the utility of the variable rule approach as showing 'a great advance in accountability over the use of "free variation"' (Labov 1969: 737, fn 20).

The important point for our argument here is that early studies in the LVC tradition were situated relative to the goal of comprehensive linguistic accountability: providing a description (either categorical or probabilistic) of *all* occurrences of variants in the targeted linguistic system. In contrast, recent variationist studies have adopted the more restricted goal of accounting for a linguistic variable – i.e. the interchangeable tokens that 'say the same thing' or are 'functionally equivalent'. Although a few recent studies have advocated consideration of non-variable linguistic tokens (see, e.g., Aaron 2010; Brook 2018), most studies focus on only interchangeable tokens while largely disregarding any obligation to account for the linguistic distribution of all variants in all contexts.

This trend has been apparent in many variationist studies of the English genitive, which have usually focused on analysis of two genitive variants (*'s* versus *of* phrases). Published accounts often include linguistic descriptions of contexts that are categorically associated with one or the other variant (see, e.g., Hinrichs & Szmrecsanyi 2007; Jankowski & Tagliamonte 2014; Grafmiller 2014). For example, *'s*-genitives are categorically required in certain conventionalized expressions and other idiomatic units (e.g. *Murphy's Law, bird's nest*). In contrast, the *of*-genitive is categorically required when the head noun is indefinite (e.g. *some members of his cabinet*), measure expressions (e.g. *a pound of flesh*) and other conventionalized expressions (e.g. *The President of the United States*).

However, many other genitive tokens are judged to be non-interchangeable – and therefore not eligible for inclusion in a variationist study – even though they cannot be accounted for on categorical grounds. Thus, coders have been required to rely on their own intuitions to exclude all tokens that are judged to be non-interchangeable. Examples (1)–(3) illustrate *of*-genitives that are likely to be coded as non-interchangeable, even though there is no clear categorical rule that states why the *'s*-genitive is not possible:

(1) in the critical areas of energy and high technology
(2) oversight of Chalabi's information operation
(3) at the scene of the latest Arab-Israeli fighting

Judging from the results reported in Hinrichs & Szmrecsanyi (2007: 445), a very large number of genitive tokens have been excluded from consideration on this basis. For

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

example, roughly 70 percent of all tokens of the *of*-genitive were excluded from the 2007 study because they were judged to be non-interchangeable. The main point for our purposes here is that variationist studies of the genitive have generally disregarded the goal of comprehensive accountability, because non-interchangeable tokens are excluded from consideration, even though it is not possible to provide a categorical explanation of why only one variant is possible for those tokens.

More recently, though, some variationist genitive studies have shifted to the broader notion of accountability advocated here, attempting to provide a fully comprehensive account of the factors influencing the choice between *'s* and *of*. That is, following the recommendations of Rosenbach (2014: 224–5), recent variationist studies describe the contexts that categorically predict the use of *'s* or *of*, and then conduct statistical analyses to probabilistically account for *all* other tokens (rather than only tokens that are deemed to be interchangeable; see, e.g., Heller *et al.* 2017). Szmrecsanyi *et al.* (2016) similarly point out the need for such research:

> Let us conclude with some philosophical afterthoughts: … How should we define the scope of a linguistic variable? And what is the appropriate domain of inquiry for studies of linguistic variation? Traditionally, the domain of inquiry for studies of genitive variation has been restricted to those linguistic tokens that are interchangeable between the *s*-genitive and the *of*-genitive. The linguistic status of noninterchangeable tokens is not clear; they have simply been disregarded as not relevant to the domain of inquiry. But this is another way of saying that at present, we know little about another type of variability: why some tokens are interchangeable and others are not – which is yet another issue that future research should address. (Szmrecsanyi *et al.* 2016: 26)

It is this broader goal of accounting for all tokens of genitive variants that motivates the present study.

In addition, a second aspect of broader accountability has also been especially relevant to the study of the English genitive: what set of variants to include in the analysis? Most early studies are based on a description of two variants: *'s* and *of*. However, Szmrecsanyi *et al.* (2016) argue that the accountability of genitive studies should be extended to include analysis of a third major variant: premodifying nouns (referred to as the NN variant below). One argument for inclusion of the NN variant is the fact that it is often interchangeable with at least one of the other two variants, often with the meaning of possession (see also Rosenbach (2014: 222), who notes 'the possessive meaning of noun modifiers'). For example, compare (4)–(6):

(4) *'s* variant: The military's predilection for extreme caution
(5) NN variant: The military predilection for extreme caution
(6) *of* variant: The predilection of the military for extreme caution

Deciding what variants to include depends on the theoretical definition of the *genitive* in English (see also Szmrecsanyi *et al.* 2016: 26). Oddly, though, none of the published papers that we reviewed include such a definition. Rather, the *genitive* is simply assumed to exist, and usually assumed to consist of two variants (*'s* and *of*). That is, authors have

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

disregarded the need for a grammatical or syntactic definition of what the genitive *is*. Rather, the underlying assumption seems to be entirely implicit: because *'s* constructions are often interchangeable with *of* constructions (and with some premodifying nouns), they must be variants of a single grammatical system.

If we instead define the English *genitive* as a linguistic system based on shared grammatical/semantic/syntactic characteristics, we could argue for one of three different analyses:

1. The term *genitive* is used to refer to a morphological case marking, and there is only one genitive construction in English (i.e. the inflectional suffix *'s*).
2. The term *genitive* is defined on semantic grounds as two noun phrases expressing the general meaning relationship of possession.
3. The term *genitive* is defined on grammatical/syntactic grounds as a noun phrase that functions as a restrictive modifier of a head noun phrase.[1]

While definition 1 adopts the meaning of the term *genitive* as it is applied across languages, it fails to capture the fact that there are other patterns of linguistic variation and choice in English that need to be accounted for. Definition 2 turns out to be unsatisfactory because only a small proportion of the tokens for all three variants express the meaning of possession. That is, even interchangeable tokens often express many other meanings, including 'prototypical' relationships, like kinship, body parts, ownership and part/whole, as well as a wide range of non-prototypical relationships, like undergoer, property, creator, theme (see Rosenbach 2014: 229; Payne & Berlage 2014; Szmrecsanyi *et al.* 2016: 26).

In contrast, definition 3 has the advantage that it describes a linguistic system in grammatical/syntactic terms. It is unconventional to refer to this system as the *genitive*. But regardless of the term that we use, definition 3 delimits a linguistic system with three main structural variants (*'s*, NN, *of*), which all function as modifying nouns that restrictively modify a head noun. Our goal in the present study is to provide a comprehensive account of variation within the linguistic system described in definition 3.[2]

Finally, we argue that the scope of accountability also needs to be extended to account for the role of register differences. The results of previous linguistic studies of register variation (see, e.g., Biber 1988, 2014; Biber & Gray 2016) lead to the strong expectation that the choice of genitive variant will be influenced by register differences. In particular, studies carried out in the TxtLx (Text-Linguistic) theoretical framework have repeatedly documented a systematic relationship between the

[1] Appositive noun phrases would be excluded from this definition because they represent a non-restrictive grammatical relationship to the head noun.

[2] Definition 3 assumes that the preposition *of* has little or no inherent semantic content, functioning instead as one of the three grammatical strategies in English for expressing the syntactic relationship of restrictive modification of a head noun. However, many other English prepositions function as restrictive noun modifiers (e.g. *in* phrases and *on* phrases), and thus might also be regarded as genitive variants. We return to this issue in the conclusion.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

situational/communicative characteristics of a register and the typical linguistic characteristics of that register (see, e.g., Biber & Conrad 2019; Biber 2019). These patterns are interpreted as instances of *functional correspondence*, built on the theoretical claim that linguistic features are frequent and pervasive in texts from a register because they are functional, serving communicative functions related to the communicative characteristics of the register (see Biber & Egbert 2023).

Recent variationist studies similarly argue for the importance of register differences as predictors of linguistic variation. For example, Szmrecsanyi (2019) notes that 'register has been shown time and again to be an important factor regulating variation' (p. 80), summarizing the results of eleven previous variationist studies. More recent studies – which have directly investigated the possibility that linguistic constraints operate to differing extents and in differing ways in different registers – support the even stronger conclusion that language users operate with different register-specific grammars (see the series of published studies on variation in the expression of future temporal reference and the dative alternation, including Engel 2022; Engel *et al.* 2022; Engel & Szmrecsanyi 2023; Szmrecsanyi & Engel 2022).

An early study in this line of research – Grafmiller (2014) – is especially relevant for our purposes here, because it focused on the alternation between the *'s*-genitive and the *of*-genitive. Grafmiller analyzed the contextual factors that influence this choice in genitive tokens from six major spoken and written registers (referred to as 'genres'), and found significantly different effect sizes for several language-internal constraints across registers. For example, possessor animacy was found to be most important as a predictor of genitive choice in the spoken register, but less important in written registers like fiction and academic writing, and the least important in newspaper writing (see Grafmiller 2014: 491).

Studies in this line of research challenge a widely accepted assumption of previous variationist research: that variation grammars are stable across styles (and registers). For example, Guy (2005: 562) notes that 'stylistic variation is quantitatively simple … indeed, it is commonly assumed in [Variable Rule] analyses that the grammar is unchanged in stylistic variation'. And Labov (2010: 265) similarly notes that 'internal constraints … are normally independent of social and stylistic factors'. In contrast, recent studies like Szmrecsanyi & Engel (2022) and Engel & Szmrecsanyi (2023) show that register differences directly influence the conditioning of grammatical variation in three major respects: the set of relevant constraints; the effect size of the constraints; and the direction of the effect. These recent studies have focused on grammatical alternations like *will* versus *be going to*, and the object versus prepositional dative. Thus, a third goal of the present study is to extend the results of those studies to the English genitive, exploring the ways and extent to which register influences the choice among genitive variants.

In sum, the present study hopes to extend the accountability of previous variationist studies of the genitive in three major ways:

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

1. It attempts to account for *all* tokens of genitive constructions: categorical, interchangeable, and non-interchangeable but non-categorical.
2. It accounts for the choice among three major genitive variants: *'s*, NN and *of*-phrases.
3. It accounts for the possibility of register differences, exploring the ways in which constraints interact with one another in different registers, as well the possibility of constraints having different importance in different registers.

2 Corpus and sample of genitive tokens

We began with a corpus that was equally matched for three registers: American English conversation (from the *Longman Spoken and Written Engish Corpus*; see Biber *et al.* 2021: 25–35), newspaper articles (from the *New York Times*), and (social) science research articles (from the *20th Century Research Article Corpus*; see Biber & Gray 2016: 52–7). The corpus was exactly balanced across the registers: 150 texts from each register, where each text was 667 words long. Thus, the entire corpus consisted of 450 texts (300,150 words), with 100,050 words of text from each register.

All occurrences of N*'s*/NN/N+*of* in the corpus were automatically extracted. We then manually checked all occurrences to delete any that had been incorrectly included. This resulted in a total pool of 14,405 genitive tokens in the corpus.

We analyzed the distribution of genitive variants from both text-linguistic and variationist perspectives. The requirements of the sample are very different in a text-linguistic study versus a variationist study. In a text-linguistic study (see section 4 below), the goal is to describe the rates of occurrence for each variant (i.e. how often we encounter a linguistic feature in a fixed amount of text), and so the analysis should be based on all occurrences of the variants in all texts. Thus, the text-linguistic description was based on analysis of the total pool of 14,405 genitive tokens in the corpus (see figure 1 in section 4). In contrast, the goal of analysis in a variationist study is to accurately predict the contextual factors that favor the use of one variant over another (i.e. what contextual factors favor the use of a variant, when it is used – regardless of the rate of occurrence for that variant in texts). In this case, we want a large random sample of tokens that represents the range of uses and grammatical contexts for each variant, but with no requirement to include all tokens from the corpus.

For the variationist analysis, we randomly selected a sample of genitive tokens from each register, and subsequently annotated each of those tokens for its contextual characteristics. Practical considerations limit the size of the sample for the variationist analysis, because each token needs to be annotated manually. Thus, the goal is to obtain a sample that is large enough to accurately represent the contextual patterns of variation, but small enough to make the hand-coding feasible. For the present study, we set a target of 3,000 genitive tokens for the variationist analysis. However, because we initially extracted genitive tokens randomly from the corpus, we ended up with larger samples of the more frequent variants (especially the NN variant). Therefore, we increased the samples for the rare variants (especially the *'s*-genitive) to represent as wide a range of variation for those variants as possible. We attempted to include at

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

least 300 tokens for each variant from each register, and if there were not 300 tokens of a variant in a register, we included all available tokens.

In previous variationist studies, the unequal representation of variants in the sample is assumed to have no bearing on the results, because the goal is to determine the factors associated with each variant, *when that variant occurs* (as opposed to describing the extent to which a variant is used in texts). So, as noted above, the essential consideration is to annotate a random sample that is as large as possible, with no requirement that the sample be balanced with the same number of tokens for each variant. (We return to a fuller discussion of this point in section 6.)

The sample of genitive tokens for the variationist analysis was pre-edited to exclude all cases that could be accounted for on categorical grounds, that is, cases where only one variant was possible because of a categorical rule:

Only *'s* is possible:

- *'s* with no head noun (e.g. *it was Sam's*)

Only *of* is possible:

- pronominal modifiers (e.g. *a piece of that*, *that picture of me*)
- idioms (e.g. *a matter of fact*)
- quantifiers that are not nouns (e.g. *much of*, *several of*, [x]*%* + *of*)
- idiomatic quantifiers (e.g. *a lot of*, *lots of*, *little bit of*)
- numbers (e.g. *billions of dollars*)
- constructions with a complement clause filling the NP slot (e.g. *a question of what I want to do*)

Only NN is possible:

- title nouns (e.g. *President Biden*)
- proper nouns (e.g. *Pennsylvania Steel Company*)
- appositives (e.g. *fitness expert Dr Jones*)

In addition, we excluded constructions with a possessive determiner and constructions that could be interpreted as having an adjectival premodifying structure:

- possessive determiners (e.g. *her trip*, *my excuse*)
- modifiers consisting of a compound participle (e.g. *gas-sipping autos*, *state-owned business*)
- noun/adjective ambiguities referring to nationalities or groups (e.g. *American*, *Chinese*, *Christian)*

It is worth emphasizing that this step in the analysis did not require intuitive judgments of interchangeability. Rather, all tokens in the sample that could not be clearly accounted for by a categorical rule were included in the subsequent statistical analysis, regardless of their interchangeability.

Using the sampling methods outlined above, and after exclusion of the categorical tokens, we obtained a sample of 3,425 genitive tokens (see table 1). These tokens were

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Table 1. *Number of annotated genitive tokens for each variant in each register*

	Academic writing	News articles	Conversation	TOTAL
's	153	303	80	536
NN	861	493	601	1,955
of	384	299	251	934
TOTAL	1,398	1,095	932	3,425

annotated for the contextual characteristics described in section 3. Just to be clear, even though this sample does not include genitive tokens that can be accounted for on categorical grounds, there are many other tokens included in the analysis that might be judged to be non-interchangeable. Thus, the present study attempts to achieve greater accountability by analyzing the factors that predict these non-variable tokens that would have been disregarded in previous analyses.

3 Linguistic annotation of genitive tokens

For our analysis here, we refer to the two grammatical components of a genitive construction as the *Head noun* and the *Modifying noun (phrase)*. This departure from previous practice, which has usually employed the terms *possessor* and *possessum*, gives primacy to the syntactic functions of the two components, and avoids any suggestion that the core meaning of the genitive construction in English relates to 'possession'.

Each of the 3,425 tokens in our sample was coded for its genitive variant (*'s*, NN, *of*) and eleven contextual characteristics: register; length of the Modifying NP; semantic category of the Modifying noun (i.e. the head noun of the modifying NP); semantic category of the Head noun (i.e. the head noun of the head NP); final sibilancy of the Modifying noun; Modifying noun to Head noun thematicity ratio; and five specific structural/syntactic characteristics of the genitive construction (other premodifiers in the Head NP, other postmodifiers in the Head NP, postmodifiers in the Modifying NP, the embedding of the Head NP in a higher-level PP, and the embedding of the Head NP in a higher-level NP). Detailed descriptions follow:

Register: Each token was coded for register: conversation, newspaper article, academic research article.

Length of the Modifying NP (mod_length): This is the length of the modifying noun phrase in words (excluding determiners). For example:

(7) [student contact] time — mod_length = 2
(8) the presence of [clear and repeating list markers] — mod_length = 5

Semantic category of the head noun of the Modifying NP (mod_semantic); Semantic category of the head noun of the Head NP (head_semantic)

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Probably the best-established finding from previous research on genitive variation is the strong effect of the animacy of the modifying noun as a predictor that favors the *'s*-genitive (see discussion in Rosenbach 2014, especially p. 231). However, studies like Rosenbach (2002), Payne & Berlage (2014) and Egbert & Davies (2019) have documented the existence of a wide range of other semantic relationships between the Head NP and the Modifying NP in genitive constructions.

Building on this previous research, we developed a coding framework that could be used to annotate the semantic relations between Head noun and Modifying noun. That framework was modified and tested through four major rounds of piloting, resulting in a framework with fourteen major semantic relationships (e.g. composition, location/time, partitive, category, co-reference, purpose, process/experience). However, we failed to achieve acceptable reliability among coders in any of the rounds of pilot testing, while at the same time, we uncovered additional semantic relationships during each round of piloting.

As a result, we employed an alternative approach in the present study: we developed a framework with seven major semantic categories and then coded the category of the Head noun and the Modifying noun separately, with no attempt to identify the specific semantic relationship between the two. Thus, all Head nouns and all Modifying nouns in the sample were annotated as belonging to one of the following semantic categories:[3]

- Animate [A]: operationalized if the noun has agency
- Group/Institution [G]: a specific group, institution, community. If a group noun has agency, it was still coded as G.
- Location/Time [L]: a specific place or time (excluding duration or amount)
- Concrete/Tangible [C]: tangible objects/entities that are not animate
- Abstract [X]: ideas, processes, entities that are not tangible
- Quantity/Amount [Q]: specifies an amount or quantity, including size or duration (e.g. *bit*, *chunk*, *grain*, *piece*, *collection*, *set*, *part*, *unit*)
- Proper name (corporations) [P]: this code was rarely used, usually for modifying nouns that identified a corporation or agency (e.g. *NBC*, *UNOCAL*) (if the head of the head NP was a proper name, the token was excluded from analysis)

This approach proved to be much more feasible and reliable (with agreement rates between coders greater than 90 percent).

Final sibilancy of the Modifying noun (mod_sibilant): a dichotomous variable coding whether the modifying noun ends in a sibilant (*s*, *z*, *sh*, *zh*)

Modifying noun to Head noun thematicity ratio: to compute this variable, we began with analysis of the normalized rates of occurrence (per 1,000 words of text) of each

[3] We also explored using the concatenation of the Modifying+Head semantic categories as a variable, to operationalize the semantic relationship between the two noun phrases. However, the statistical results using this variable were very difficult to interpret, because there are so many possible combinations. In addition, inclusion of the concatenated variable resulted in very little improvement in predictive accuracy over inclusion of the two independent variables. Thus, the final analyses employed only the two independent semantic variables.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Modifying noun and each Head noun, in each text. Those rates were then combined into a single ratio, based on the hypothesis that it is the relative 'thematicity' of those two nouns that influences the choice of premodifying versus postmodifying genitive variant. In general, Modifying nouns had much higher rates of occurrence than Head nouns.

Specific structural characteristics

Specific structural characteristics of the Head NP and Modifying NP were coded, based on the hypothesis that simpler noun phrases are more likely to be premodifiers, and more complex noun phrases are more likely to be postmodifiers.

a. **Other premodifiers in the Head NP? (head_w_premod):** codes the existence of one-or-more adjectives or other nouns as premodifiers in the head NP (i.e. beyond the targeted genitive form)
b. **Other postmodifiers in the Head NP? (head_w_postmod):** codes the existence of one or more postmodifiers in the head NP (i.e. beyond the targeted genitive form)
c. **Head NP embedded in a higher-level PP? (head_in_pp):** y/n
d. **Head NP embedded in a higher-level NP? (head_in_np):** y/n
e. **Other postmodifiers in the Modifying NP? (mod_w_postmod):** codes the existence of one or more postmodifiers in the Modifying NP

4 Descriptive patterns of variation

4.1 Text-linguistic rates of occurrence across registers

As background to our variationist study, we undertook a text-linguistic analysis to compare the rates of occurrence of each variant in texts from each register. This perspective on language use – reflecting the extent to which texts from these registers use these grammatical devices – complements the variationist perspective, which focuses on the factors that influence the choice among the variants (regardless of their textual frequencies).

Figure 1 shows that all three genitive variants occur more frequently in the written registers than in conversation. The *'s*-genitive is by far most common in news reportage, both in terms of the rate of occurrence (817 in the 300,000-word corpus, or about 3 occurrences in a typical 1,000-word text) and proportionally (accounting for 29.2 percent of the total genitive tokens). In conversation, the *'s*-genitive is infrequent and surprisingly accounts for only 7 percent of the total genitives. In academic writing, the *'s*-genitive accounts for only 2 percent of all genitives, but it actually occurs in texts more frequently than in conversation.

NN genitives are especially common in written academic texts (with about 15 occurrences in a typical 1,000-word text), but they are also very common in news reportage. Proportionally, NN constructions account for the majority of genitive tokens in all three registers (66.8% in conversation, 54% in news, 55.2% in academic). Finally, *of*-genitives are by far the most common in academic writing, and moderately common in news. Proportionally in all three registers, *of*-genitives account for a

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

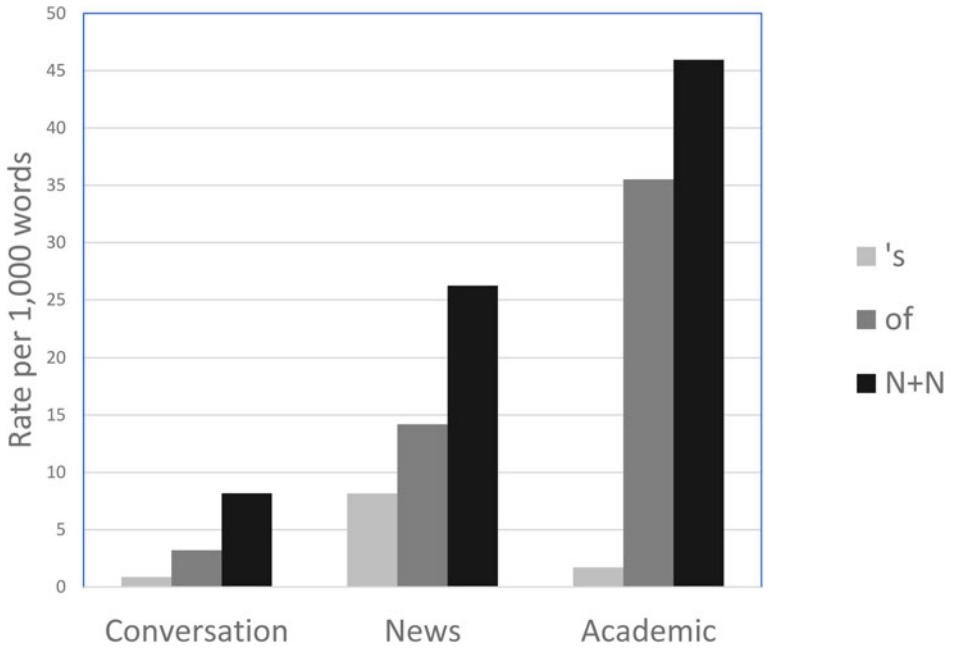


Figure 1. Rates of occurrence of three genitive features across registers

considerably larger percentage of total genitives than *'s*-genitives, but a smaller percentage than NN genitives (i.e. 26.1% in conversation, 29.2% in news, 42.7% in academic).

4.2 Descriptive patterns for the variants across contexts of use

Descriptive statistics for the predictor variables show interesting differences in the typical contextual characteristics of the three genitive variants. In addition, there are often quite different patterns of use across the three registers. For example, table 2 shows that the modifying noun phrase is generally longer in *of*-genitives than in *'s* or NN genitives. However, this difference is much more evident in academic writing than in the other two registers. In fact, 29 percent of the *of*-phrase genitives in academic writing contain a noun phrase five words or longer, as in (9):

Table 2. *Length of the modifying NP (in words)*

	Conversation		News		Academic	
	Mean	SD	Mean	SD	Mean	SD
NN	1.14	0.42	1.20	0.44	1.20	0.43
of	1.62	0.73	2.52	1.90	3.85	2.85
's	1.10	0.34	1.14	0.38	1.28	0.60

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

(9) A focal interest in seriality marked […] a fusion of the enduring issues of association of ideas within philosophy with the methods of science.

The semantic categories of Head nouns and Modifying nouns also show interesting associations with the different variants interacting with register differences. Table 3 shows that Head nouns in the three registers tend to come from a range of different semantic categories: Concrete nouns in conversation (e.g. *fruit salad*, *Jennifer's house*); Abstract nouns in academic writing (e.g. *ability parameter*); and a mix of Animate nouns, Concrete nouns and Abstract nouns in news (e.g. *Dean's campaign chairman*, *Iraq's wetlands*, *business style*). However, there are some interesting interactions with the choice of variant. For example, in conversation, Locative Head nouns and Abstract Head nouns occur with notably high proportions of the *of* variant (e.g. *middle of everything*, *sense of responsibility*).

There are stronger differences across the variants in their associations with the semantic category of the Modifying noun (table 4). The *'s*-genitive is strongly associated with Animate Modifying nouns, especially in conversation and academic writing (e.g. *my mother's dress*, *student's proficiency*). In conversation, the other two variants (*of* and NN) often occur with Concrete Modifying nouns (e.g. *the middle of the table*, *Chinese finger trap*), although a surprisingly high proportion of these tokens occur with Abstract Modifying nouns (e.g. *concept of self-accord*, *audio tape*). Academic writing shows the opposite pattern: most *of*-genitives and NN genitives occur with Abstract Modifying nouns (e.g. *admissions criteria*, *validity of teacher's assessments*), but a surprisingly high proportion of those variants occur with Concrete Modifying nouns (especially for NN constructions; e.g. *clock face array*). And here again, News shows a reliance on a wider range of different semantic categories. The *'s*-genitive is the most distinctive pattern in News (in contrast to the other registers): less than half of these *'s*-genitives occur with Animate Modifying nouns, while relatively large proportions occur with Group and Locative Modifying nouns (e.g. *the Army's parachute team, Boston's Logan Airport*). In addition, Concrete nouns, Group nouns, Locative nouns, and Abstract nouns are all relatively common as Modifying nouns with the *of* and NN variants in News.

Previous research has documented the influence of a final sibilant (*s*, *z*, *sh*, *zh*) on the Modifying noun as a factor favoring the *of*-genitive over the *'s*-genitive. Table 5 shows that *'s*-genitives are rarely used when the Modifying noun ends in a sibilant, especially in conversation. This tendency is less strong in writing. For example, in academic writing we find 10.5 percent of all *'s*-genitives occur with a Modifying noun that ends in a sibilant (e.g. *the task force's prosecutors*). NN genitives also rarely occur when the Modifying noun ends in a sibilant (in all three registers), while *of*-phrases occur much more freely in this context (especially in the two written registers).

Table 6 shows that Modifying nouns generally have higher thematicity (i.e. higher rates of occurrence in the text) than Head nouns, for all three variants. However, thematicity has a somewhat different pattern in the three registers. In conversation, Modifying nouns with

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Table 3. *Semantic category of the Head noun (head_semantic) (percentage)*

	Animate	Concrete / Tangible	Group / Institution	Location / Time	Proper name	Quantity / Amount	Abstract	TOTAL
Conversation								
NN	8.5	71.2	1.3	0.0	0.0	0.0	19.0	100
of	5.1	44.1	0.7	14.9	0.0	4.4	30.8	100
's	10.0	77.4	0.0	1.3	0.0	0.0	11.3	100
News								
NN	23.5	34.5	13.4	2.2	0.6	0.0	25.8	100
of	16.9	27.2	5.0	4.0	0.0	8.6	38.2	100
's	23.1	33.3	8.3	1.0	0.0	0.3	34.0	100
Academic								
NN	5.5	12.0	2.3	0.2	0.0	0.3	79.7	100
of	0.5	5.5	2.6	2.6	0.0	8.9	79.9	100
's	2.0	13.1	0.0	0.7	0.0	0.0	84.3	100

Table 4. *Semantic category of the Modifying noun (mod_semantic) (percentage)*

	Animate	Concrete / Tangible	Group / Institution	Location / Time	Proper name	Quantity / Amount	Abstract	TOTAL
Conversation								
NN	2.0	64.9	1.7	8.3	4.2	1.8	17.1	100
of	12.2	56.3	2.4	5.8	1.0	0.0	22.4	100
's	83.8	0.0	2.5	12.5	0.0	0.0	1.3	100
News								
NN	2.8	29.6	15.2	11.4	7.1	0.8	33.1	100
of	14.0	33.6	10.3	10.6	2.7	1.3	27.6	100
's	44.9	5.6	26.7	20.8	0.0	0.0	2.0	100
Academic								
NN	8.9	21.0	2.4	3.0	3.8	0.6	60.2	100
of	8.3	10.9	1.6	2.9	0.3	0.5	75.5	100
's	82.4	6.5	3.9	2.0	0.0	1.3	3.9	100

Table 5. *Percentage of the variant (in each register) where the Modifying noun ends in a sibilant (mod_sibilant)*

	Conversation		News		Academic	
	No	Yes	No	Yes	No	Yes
NN	85.4	14.6	84.8	15.2	82.7	17.3
of	78.0	22.0	64.5	35.5	53.6	46.4
's	95.0	5.0	92.4	7.6	89.5	10.5

all three genitive variants tend to be more thematic than the associated Head noun, but this pattern becomes extreme in the case of *of*-genitives. In contrast, there are only small differences across the variants in news, with *'s*-genitives exhibiting the strongest tendency to employ a Modifying noun with higher thematicity than the Head noun. Academic writing follows the same general pattern as conversation, with the Modifying noun in *of*-phrases having the highest thematicity values in comparison to the associated Head noun thematicity values. However, there are relatively small differences in thematicity ratio across the three variants in adacemic writing, with the general pattern being that Modifying nouns tend to have much higher thematicity than Head nouns in all genitive constructions.

We coded for the existence of additional modifiers in the Head noun phrase, based on the hypothesis that an additional premodifier might favor the use of a postmodifying genitive variant, while an additional postmodifier might favor the use of a premodifying genitive variant. In conversation, the existence of an additional premodifier in the Head NP favors an *of*-genitive over the NN or *'s* variants (see table 7). But the existence of an additional premodifier has only a weak relationship to the choice of genitive variant in the two written registers. In contrast, the existence of an additional postmodifier in the Head NP (table 8) has a much stronger relation to genitive choice in writing than in speech, with the *of*-genitive being dispreferred relative to the NN and *'s* variants in that context.

We also coded for the case when the genitive construction was itself embedded in a higher-level prepositional phrase (table 9) or higher-level noun phrase (table 10), with

Table 6. *Ratio of Modifying noun thematicity to Head noun thematicity*

	Conversation		News		Academic	
	Mean	STD	Mean	STD	Mean	STD
NN	1.92	3.09	3.01	4.79	7.17	23.56
of	17.17	48.98	2.61	3.62	12.69	27.09
's	2.64	3.72	5.81	6.53	10.24	21.60

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Table 7. *Percentage of the variant (in each register) where the Head NP has an additional premodifier (head_w_premod)*

	Conversation		News		Academic	
	No	Yes	No	Yes	No	Yes
NN	83.9	16.1	69.8	30.2	63.8	36.2
of	74.9	25.1	63.8	36.2	65.4	34.6
's	93.8	6.3	56.8	43.2	64.7	35.3

Table 8. *Percentage of the variant (in each register) where the Head NP has an additional postmodifier (head_w_postmod)*

	Conversation		News		Academic	
	No	Yes	No	Yes	No	Yes
NN	93.0	7.0	70.4	29.6	77.0	23.0
of	96.9	3.1	98.0	2.0	95.6	4.4
's	95.0	5.0	70.0	30.0	62.7	37.3

the hypothesis that such embedding would favor the choice of a more compressed genitive variant (i.e. either the NN or *'s* variant). Table 10 indicates that this pattern holds to some extent in the written registers, favoring the use of the NN variant when the Head NP is embedded in a higher-level NP. The results in table 9 are less clear, because all three variants seem to occur freely when the Head NP is embedded in a higher-level PP; however, there appears to be a slight trend in this context favoring the premodifying genitive constructions in academic writing.

Finally, we analyzed the influence of additional postmodifiers in the Modifying NP, based on the hypothesis that such structural elaboration would be hard to process if it occurred in a premodifying genitive NP. Table 11 shows that hypothesis is strongly

Table 9. *Percentage of the variant (in each register) where the Head NP is embedded in a higher-level PP (head_in_pp)*

	Conversation		News		Academic	
	No	Yes	No	Yes	No	Yes
NN	76.0	24.0	63.7	36.3	49.8	50.2
of	80.7	19.3	64.5	35.5	62.2	37.8
's	78.8	21.2	61.4	38.6	47.7	52.3

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Table 10. *Percentage of the variant (in each register) where the Head NP is embedded in a higher-level NP (head_in_np)*

	Conversation		News		Academic	
	No	Yes	No	Yes	No	Yes
NN	95.5	4.5	90.1	9.9	87.0	13.0
of	99.9	0.1	100.0	0.0	100.0	0.0
's	100.0	0.0	100.0	0.0	99.3	0.7

Table 11. *Percentage of the variant (in each register) where the Modifying NP has an additional postmodifier (mod_w_postmod)*

	Conversation		News		Academic	
	No	Yes	No	Yes	No	Yes
NN	100.0	0.0	99.9	0.1	100.0	0.0
of	97.3	2.7	94.0	6.0	67.7	32.3
's	100.0	0.0	100.0	0.0	100.0	0.0

supported. In particular, academic writing strongly exploits this difference between the premodifying and postmodifying genitives, with nearly one-third of all *of*-genitives occurring with an additional postmodifier.

5 Accounting for the choice of genitive variant

Section 4.2 presents descriptive information showing that the three genitive variants are to a large extent associated with different grammatical/semantic contexts, with patterns that vary in systematic ways across registers. However, such univariate descriptions cannot sufficiently account for the choice among variants: we do not know the overall extent to which we can predict genitive choice when all predictors are considered; we do not know the conditional importance of each linguistic predictor (i.e. the extent to which a predictor accounts for genitive choice when all other predictors are also taken into account); and we do not know the specific ways in which predictors interact with one another. To achieve these goals, we employ two complementary statistical techniques: conditional inference trees and random forests (see Tagliamonte & Baayen 2012; Szmrecsanyi *et al.* 2016; Fahy *et al.* 2022).

Conditional inference trees are useful because they make it easy to visualize how multiple predictors interact with one another. The conditional inference tree algorithm recursively splits predictor values in a way that results in the 'purest' possible subgroups of observations. In our case, this means that the analysis uses contextual

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

factors to attempt to create groupings consisting of tokens from a single genitive variant. The result is visualized as a flowchart-like decision tree (see examples below).

Although conditional inference trees are easy to interpret, they can be somewhat unstable across different datasets, because each split in the tree depends on the splits that precede it. The random forest approach was developed to address this problem, based on analysis of a 'forest' of conditional inference trees. Each tree is restricted to a random sample of observations and a random subset of predictors. By pooling the results over the entire 'forest' of trees, the approach produces a model that is highly robust. A random forest analysis provides information about the relative importance of predictors and the overall accuracy of prediction, in a way that has certain advantages over traditional regression analyses: it is minimally affected by problems of predictor multicollinearity and data overfitting, and importantly for our purposes here, it can be readily applied to the analysis of a linguistic variable with three or more variants (see, e.g., Tagliamonte & Baayen 2012).

We ran separate statistical analyses for each register, to explore the possibility that linguistic constraints operate to differing extents and in differing ways in the three different registers.[4] From the random forest analyses, we focus on two main types of output: the overall predictive accuracy of the models, and the variable importance plots, which graphically show the relative conditional importance of each predictor in the model.

Table 12 presents the prediction accuracies of the random forest models, contrasted with the predictions that would have resulted from random chance. Overall accuracy rates are similarly high for all three registers: >80 percent of all tokens are correctly predicted, contrasted with rates of 8–60 percent for predictions based on random chance. These results are directly relevant for our first research goal, demonstrating that we are generally able to account for *all* tokens of genitive constructions (including non-interchangeable tokens).

However, there are some noteworthy differences for the three variants. Table 12 presents both precision rates (i.e. the extent to which the variant predicted by the model is correct) as well as recall rates (i.e. the extent to which the actual occurrences of a variant are correctly predicted). In general, the model performs better for the precision rates of NN and *'s* variants, but better for the recall rates of the *of* variant. Precision rates are especially high for the NN variant. That is, when the model predicts that a token is an NN variant, it is usually correct. But recall rates are lower: the model fails

[4] We alternatively could have presented the results from a single random forest analysis on the entire sample of genitives, with *register* as a predictor. However, that approach is less suitable for the purpose of comparing the ways in which genitive variation is structured within each register, because:

a. *register* is a very different type of variable than all other variables included in the analysis – it relates to the external context of texts, rather than to the immediate linguistic context of tokens, and it has only three values in the present study, even though there are many more possible registers in English
b. the resulting conditional inference tree is extremely difficult to interpret, because it is extremely large, with *register* appearing repeatedly as a node that interacts with the other contextual factors.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Table 12. *Accuracy rates (precision and recall, contrasted with random chance) of the random forest models for each register (percentage)*

	Overall accuracy	NN			*of*			*'s*		
		Precision	Recall	Random chance	Precision	Recall	Random chance	Precision	Recall	Random chance
Conversation	84.1	93.3	85.0	61.6	66.4	84.1	30.2	80.0	77.1	8.2
News	80.2	83.4	81.7	44.9	69.8	85.0	27.4	85.5	74.6	27.6
Academic	85.2	92.2	85.6	61.6	71.4	91.9	27.5	76.5	70.9	10.9

to correctly account for around 15 percent of the actual NN tokens. The *of*-variant shows the opposite pattern: the high recall rates show that the model correctly accounts for 85–90 percent of all actual *of* tokens, but the lower precision rates show that around 30 percent of the tokens that are classified as *of*-genitives are actually tokens of another variant.

Figures 2–4 present variable importance plots for each register. The semantic category of the Modifying noun and the length of the Modifying NP are the two most important predictors in all registers. However, beyond that similarity, there are two noteworthy differences across registers:

- Academic writing shows a reversal of the two most important predictors, with the length of the Modifying NP being considerably more important than the semantic category of the Modifying noun.
- The semantic category of the Head noun is moderately important in Conversation (and less so in News), but not in Academic writing.

Table 13 provides a schematic synthesis of the findings across the three models (from figures 2–4), reflecting both the relative ranking of predictors as well as the extent to which a predictor is influential. In section 1, we surveyed recent studies that demonstrate the importance of register as a mediating factor in accounts of linguistic variation. The patterns summarized in table 13 confirm that influence. In particular, the organization of linguistic constraints in the three registers differ with respect to the sets of relevant constraints, the effect size of constraints, and even the relative strength of the two most important predictors. We return to a discussion of the theoretical implications of these differences in section 6 below.

The results of the random forest analyses are useful for determining the overall extent to which linguistic variation can be predicted as well as the relative importance of each predictor. However, they do not provide the kinds of detailed information required to fully account for the patterns of variation. Conditional inference trees (CITs) are more useful for these purposes, providing three kinds of information not readily available from a random forest analysis:

Table 13. *Comparative summary of the relative importance of predictors in each register*

	Conversation	News	Academic
Mod_semantic	++++++++++	++++++	++++
Mod_length	+++++	++	+++++
Head semantic	++	+	
Mod/head theme	+		+
Head w premod	+		
Head w postmod			+
Mod sibilant		+	

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

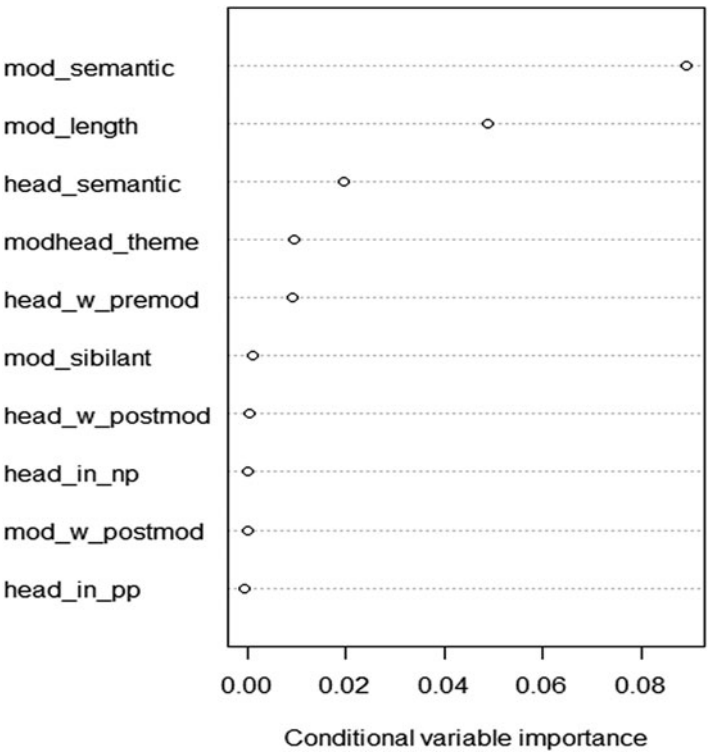


Figure 2. Variable importance plot for conversation
Key:
mod_semantic: semantic category of the Modifying noun
mod_length: length of the Modifying NP
head_semantic: semantic category of the Head noun
modhead_theme: ratio of Modifying noun thematicity to Head noun thematicity
head_w_premod: there is an additional premodifer in the Head NP
mod_sibilant: there is a final sibilant in the Modifying noun
head_w_postmod: there is an additional postmodifer in the Head NP
head_in_np: the Head NP is embedded in a higher-level NP
mod_w_postmod: there is an additional postmodifer in the Modifying NP
head_in_pp: the Head NP is embedded in a higher-level PP

1. A schematic representation of the specific ways in which predictors interact with one another
2. Specification of the particular linguistic variant that is favored by each configuration of predictors
3. Specification of the particular values of a predictor that favor a variant

All three of these characteristics of CITs are essential for a full interpretation of the patterns of linguistic variation. The results from the random forest analyses

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

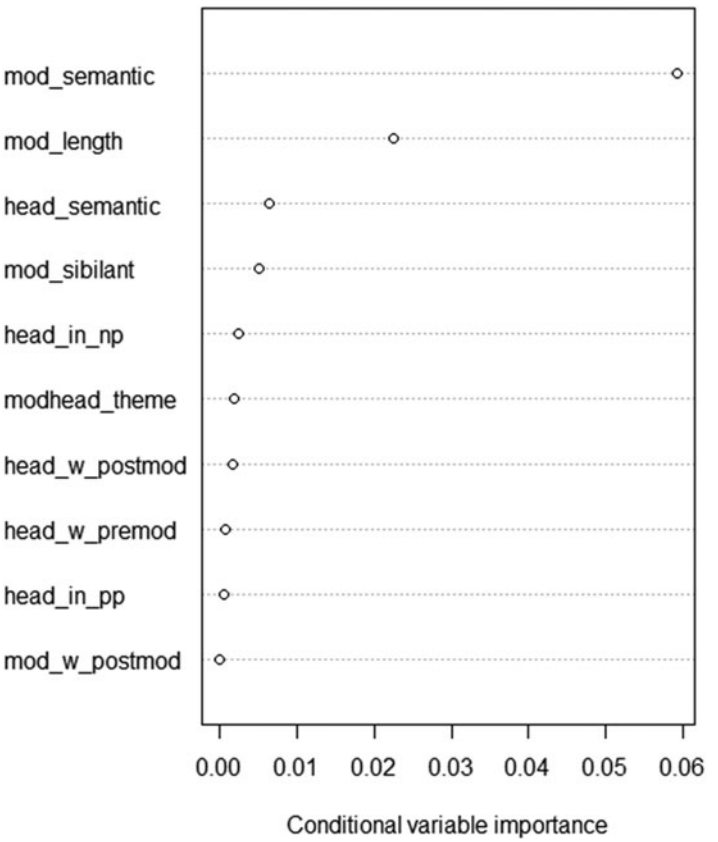


Figure 3. Variable importance plot for news articles

(figures 2–4) describe the overall conditional importance of predictors (i.e. the importance of a predictor when all other predictors are considered), but those results provide no information relating to the ways in which the predictors interact. Perhaps even more important, figures 2–4 provide no information regarding the particular linguistic variants that are being predicted by each independent variable. But there is a third complication that is equally important: most of these predictor variables have many different values. For example, the semantic category of the Modifying noun (mod_semantic) has seven values (see table 4 above). Figures 2–4 show that this variable is important in all registers, but we do not know if there are particular semantic categories that matter more than others. In fact, as we'll see below, a single predictor variable can be influential at multiple points in the prediction of a linguistic variable, with different values of the predictor favoring different variants.

Figures 5–7 present the conditional inference trees (CITs) for each register. For binary variables, the accuracy of a CIT analysis can be evaluated using the Index of Concordance *C*, but there are no comparable methods for directly evaluating the accuracy of a model for

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

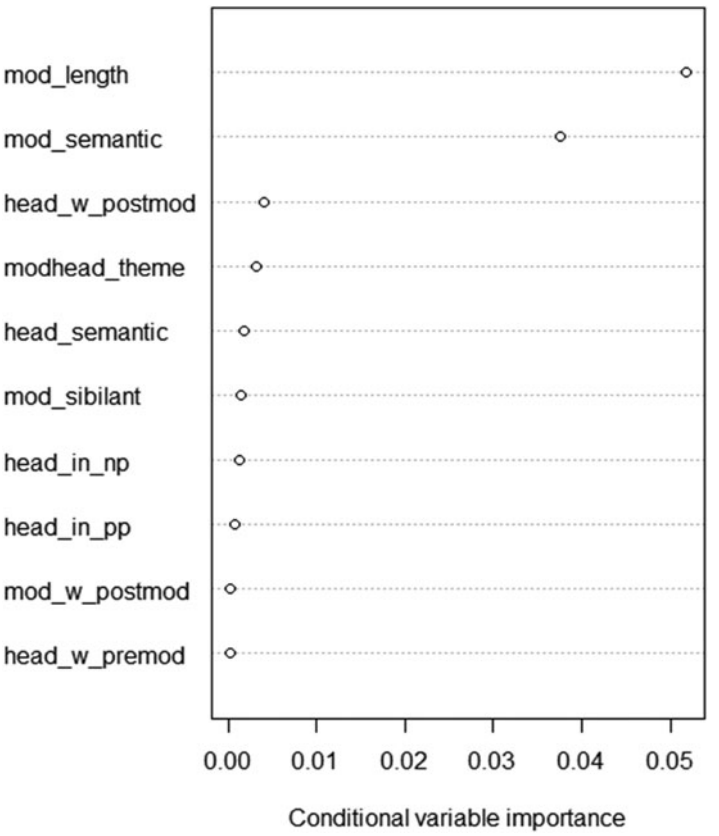


Figure 4. Variable importance plot for academic research articles

a ternary variable. So, to obtain a rough estimate of the goodness-of-fit of our approach, we calculated *C* scores for three CITs with binary response variables (*'s* vs. *OF*, *'s* vs. NN, NN vs. *OF*). The C scores of these trees range between .89 and .97, indicating very good to excellent fit (see Tagliamonte & Baayen 2012: 156). Therefore, similar to the accuracy results from the random forest analyses, the CIT results indicate that it is indeed feasible to account for all tokens of genitive variants.

The interpretation of each conditional inference tree is straightforward: each node of the tree represents a binary split in the observations, with the goal of making each of the two groupings as 'pure' as possible (where a 'pure' group would include only tokens of a single variant). Nodes higher in a tree are more important than lower nodes (although a single predictor might appear multiple times in a tree, making its overall importance greater – see below).

The bar plots at the bottom of each branch show the proportions of each variant that have been classified into each group by the model. In addition, the bar plots specify the number of tokens in the group, giving some indication of the relative importance of this combination of contextual values.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

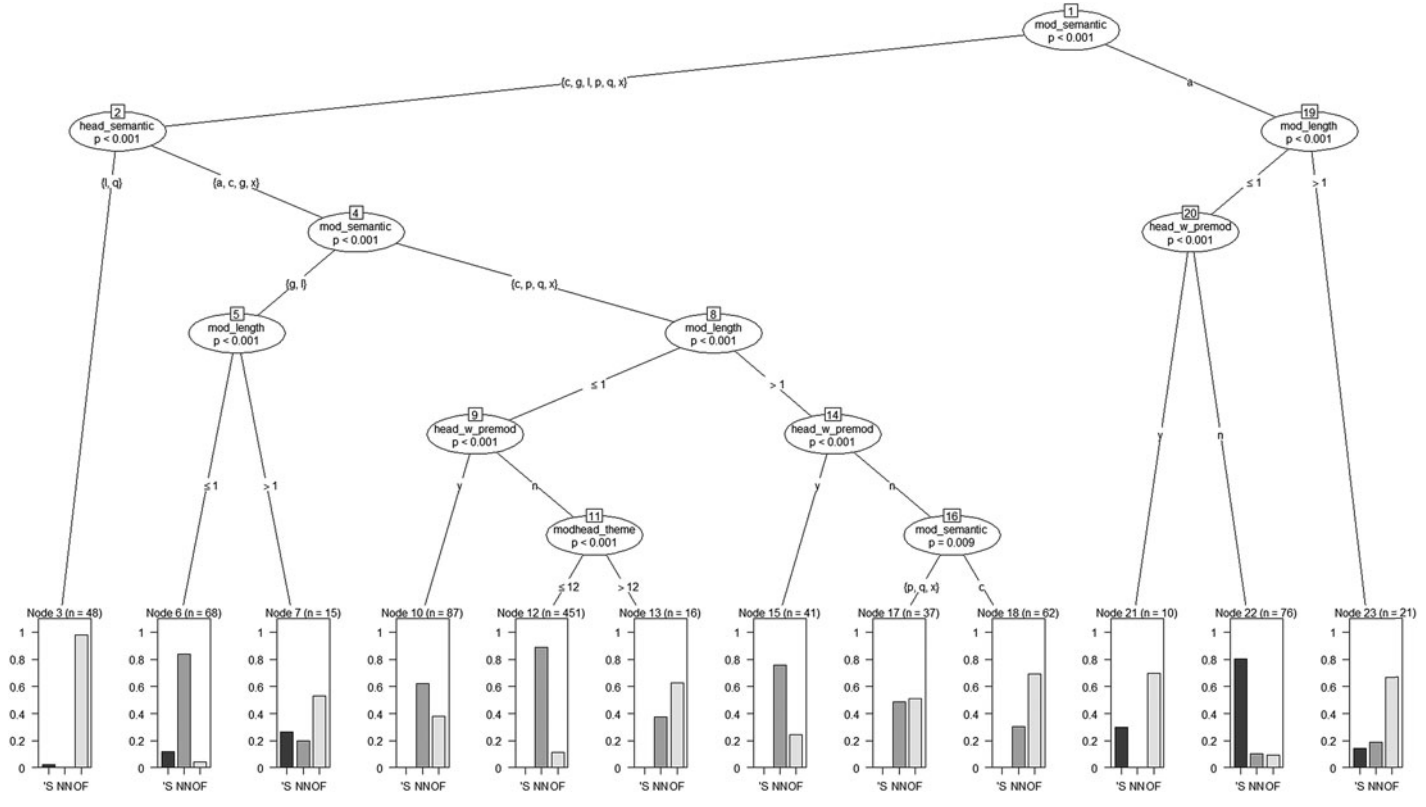


Figure 5. Conditional inference tree for conversation

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

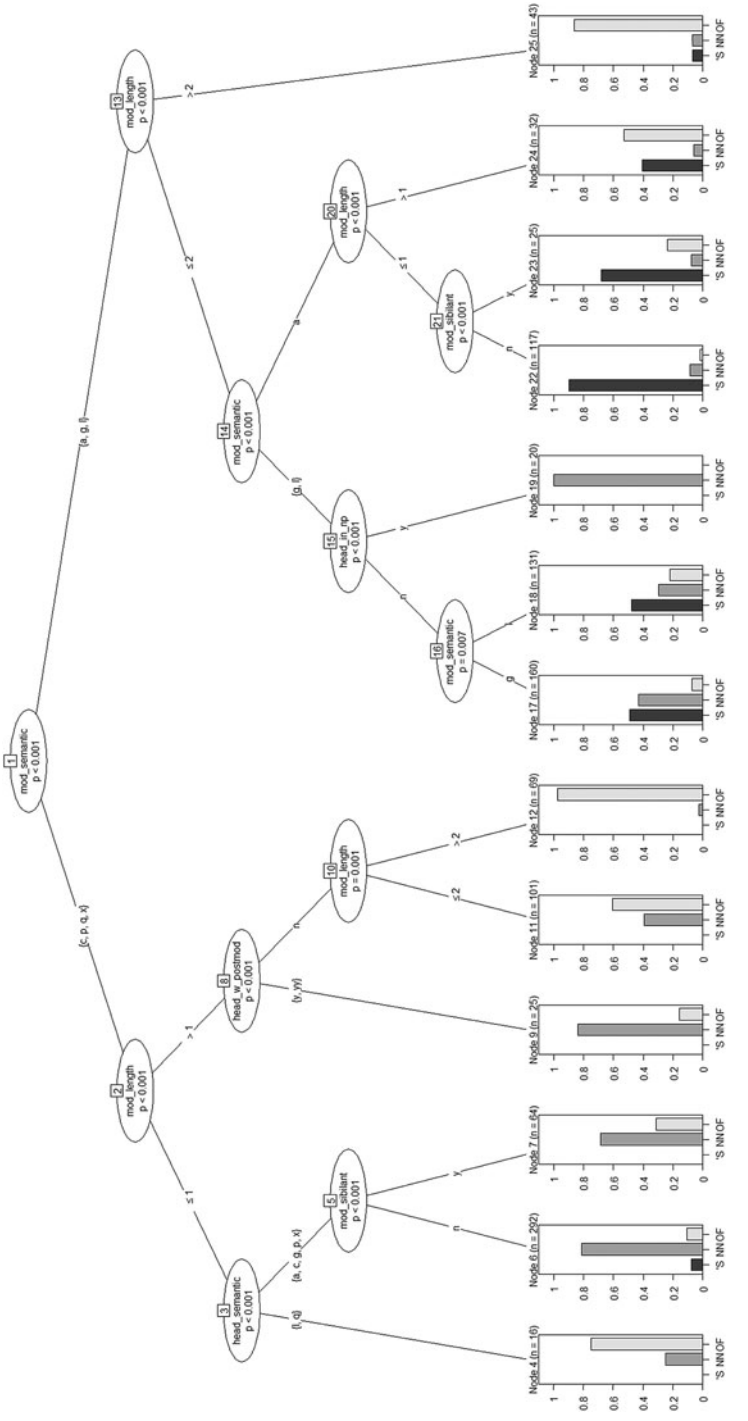


Figure 6. Conditional inference tree for news

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

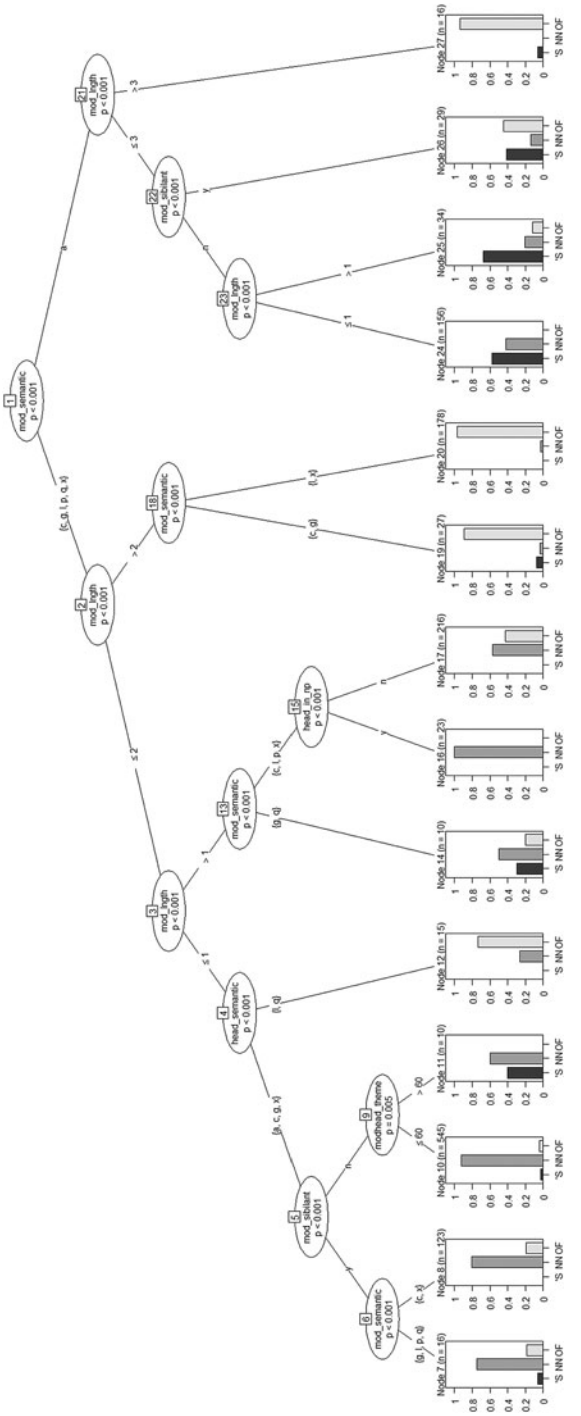


Figure 7. Conditional inference tree for academic writing

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

A visual inspection of those bar plots shows that some combinations of predictive contexts are more diagnostic than other combinations. For example, the right-most bar plot in figure 6 (for News) shows a relatively pure grouping that consists mostly of tokens with *of*. The grouping has a total of 43 tokens. Tracing the values of the branches from the top of the tree, we can see that these tokens all occur in the context of:

Mod_semantic = Animate, Group, Location
Mod_length > 2

Notice how the same predictor variable can appear at multiple points in the tree, with different splits based on different values of the variable. For example, the following branches result in a pure (but smaller) grouping that consists of only NN tokens:

Mod_semantic = Animate, Group, Location
Mod_length <= 2
Mod_semantic = Group, Location
Head_in_np = yes

In contrast, some other combinations of contextual values are less successful in specifying a pure grouping. For example, the following combination of contexts in figure 6 results in a relatively large grouping (N=131), with around 50 percent *'s*-genitives, around 30 percent NN genitives and around 20 percent *of*-genitives:

Mod_semantic = Animate, Group, Location
Mod_length <= 2
Mod_semantic = Group, Location
Head_in_np = no
Mod_semantic = Location

One interesting similarity across the CITs for all three registers is the presence of values for mod_semantic as the top-most node on the tree. In Conversation and Academic writing, this node represents the binary split between animate Modifying nouns versus all other semantic categories. The pattern is somewhat different in News, with animate, group, and locative Modifying nouns being split from the other categories (and then animate Modifying nouns being split off at a lower node).

It is important to understand how the information provided in random forest variable importance plots (figures 2–4) differs from the information provided in the CITs (figures 5–7), because the two can appear to be contradictory. Figures 2–4 plot the overall importance of each *predictor variable*. In contrast, the CITs show the importance of binary splits depending on *particular values* of predictors. The difference between the two is most apparent in the case of academic writing (compare figure 4 and figure 7). Figure 7 shows that the binary split between animate Modifying nouns versus all other semantic categories of Modifying noun is the most important contextual factor for distinguishing among the variants in academic writing. However, figure 4 shows that the predictor variable of Mod_semantic is less important than the predictor of Mod_length in this register. This apparent discrepancy can be resolved by a detailed inspection of figure 7: the distinction between animate versus all other values for mod_semantic is the

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

top-most node, but binary distinctions among other mod_semantic categories have less influence. In contrast, even for the tokens with an animate Modifying noun (the major right branch of the tree), the length of the Modifying NP turns out to be the major determinant of variant:

mod_length > 3 : favors *of*
mod_length <= 3 combined with mod_sibilant = yes: favors either *'s* or *of*
mod_length = 2–3 combined with mod_sibilant = no: favors *'s*
mod_length = 1 combined with mod_sibilant = no: favors either *'s* or NN

5.1 Describing the linguistic patterns of variation

Figures 5–7 provide a wealth of information about the ways in which contextual factors interact in the different registers to favor a particular genitive variant. While we do not have space here for a full discussion of all interactions, we can illustrate and interpret some of the major patterns.

The random forest analysis shows that the semantic category of the ModN and the length of the ModNP are the two most important predictors in all three registers. The CITs (figures 5–7) and the descriptive analyses in section 4.2 provide more details about the role of those predictors, showing that:

1. *'s*-genitives occur primarily with animate Modifying nouns
2. long Modifying NPs (even >2 words) favor the *of*-genitive, especially in the two written registers

Beyond those two general patterns, each genitive variant is favored by different interactions among the predictor variables in each register.

A. Conversation

In conversation (see figure 5), most *'s*-genitives occur with animate Modifying nouns. The *'s*-genitive is especially favored when the modifying NP length = 1 and there is no other premodifier in the Head NP, as in (10)-(11):

(10) people's motives
(11) Tracey's birthday

However, certain contextual factors favor other variants, even with an animate Modifying noun. For example, when the ModNP length is even moderately long (> 1 word), the *of*-genitive is preferred; for example:

(12) a picture of my cousin Allison
(13) the home of the conversation kings

Similarly, the presence of another premodifier modifying the head noun results in the *of*-genitive being more common than the *'s*-genitive, even with an animate modifying noun and a short modifying NP; for example:

(14) ten or twenty photos of Taylor
(15) the second son of Isaiah

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

When the modifying noun is not animate in conversation, either the *of*-genitive or the NN genitive usually occurs. The *of*-genitive is strongly favored when the head noun expresses locative/temporal or quantifying meanings, as in (16)–(17):

(16) the base of her rib cage
(17) a cup of orange juice

When the head noun expresses other meanings (and the modifying noun is not animate), the length of the Modifying NP becomes a major factor: the NN variant is strongly favored when Mod NP length = 1 (e.g. *chest pains*, *business plan*, *hair dryer*), while both the NN and *of* variants are found with longer Modifying NPs.

B. News

In News (see figure 6), the combination of an animate Modifying noun plus a one-word Modifying NP strongly favors the *'s*-genitive (e.g. *Batman's nemesis*, *Lizzie's friend*). Group and locative/temporal Modifying nouns also tend to favor the *'s*-genitive, as in (18)–(20):

(18) the regime's unease
(19) Darfur's militias
(20) America's 500 largest companies

However, the other variants (especially NN) also occur with group and locative/temporal Modifying nouns:

(21) a CIA leak
(22) Administration defensiveness over her appointment
(23) the morning newspaper
(24) world powers

And when the head NP is itself embedded in a higher-level NP, the NN variant is strongly favored with group/locative/temporal Modifying nouns:

(25) [[CIA leak] scandal]
(26) [[Capitol Hill] staff member]
(27) several former [[Riyadh embassy] employees]

The strongest factor favoring the *of*-genitive in News is a long Modifying NP (> two words), regardless of the semantic category of the Modifying noun. For example:

Long animate ModNP:

(28) the breeding of his Blue Wyandotte bantam
(29) a squad of angry army reservists

Long group ModNP:

(30) an early member of the Enron Task Force
(31) chairman of the Nevada Gaming Commission

Long concrete ModNP:

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

(32) disputed waters of the East China Sea
(33) the top of her camouflage battle-dress uniform

Long abstract ModNP:

(34) remnants of a more civilized time
(35) the immediate freezing of the entire North Korean nuclear program

However, there is additionally a special factor that favors *of*-genitives in News, even with a short Modifying NP: the presence of a locative/temporal or quantifying Head noun. For example:

Locative/temporal Head noun:

(36) outskirts of Baghdad
(37) the start of an adventure
(38) the Day of Judgment

Quantifying Head noun:

(39) troves of information
(40) plenty of ammunition

Apart from this special case, the NN variant is favored in News when the Modifying NP is short and the Modifying noun is concrete or abstract; for example:

(41) food rations
(42) aircraft engines
(43) software companies
(44) election laws

In addition, the NN variant is strongly favored with concrete/abstract Modifying nouns when the head NP is modified by an additional postmodifier, even if the Modifying NP is relatively long:

(45) Homeland-security advisor [at the EPA]
(46) attention-deficit-disorder drugs [like Ritalin and Adderall]

C. Academic research articles

Although the semantic category of the Modifying noun is less important overall in academic writing than in the other registers, this factor is very important for predicting the *'s* variant: the presence of an animate Modifying noun is the strongest factor influencing the choice of the *'s* variant in this register (see figure 7), even with relatively long Modifying NPs (2–3 words):

(47) mentally retarded children's performance in special classrooms
(48) Baker and Herman's recommendation
(49) preschool child observers' classification of naturally occurring conventional transgressions

However, tokens with animate Modifying nouns and Modifying NPs longer than three words strongly favor the *of*-genitive:

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

(50) the basic notions of numerous researchers in the area of text-structure
(51) the failure of a student enrolled at a campus during spring semester

Similar to News, Academic tokens with a non-animate Modifying noun and even a moderately long Modifying NP (longer than two words) strongly favor the *of*-genitive:

(52) the abundance of potential drift food
(53) the outcome of source monitoring processes
(54) an enigmatic pattern of egg size variation

However, the choice is more complex for tokens with a non-animate Modifying noun and short Modifying NPs (1–2 words). The NN variant predominates under these circumstances; for example:

(55) context recognition
(56) mortality rate
(57) generalizability theory

But even *'s*-genitives are found with short non-animate Modifying NPs, when the Modifying noun does not end in a sibilant, and the theme of the Modifying noun is much greater than the theme of the Head noun (Mod/Head theme > 60):

(58) the test's discrimination capacity
(59) the program's benchmark assessments

Finally, similar to the other registers, *of*-genitives are favored with a time/location or quantifying Head noun, even with a one-word Modifying NP, as in (60)–(62):

(60) beginning of the second year
(61) location of the puppet
(62) large numbers of students

6 Reconciling text-linguistic and variationist analyses of genitives

A comparison of the results presented in sections 4 and 5 confirms the importance of conducting both text-linguistic analyses and variationist analyses, because the two provide complementary perspectives on the patterns of variation (see also Biber *et al.* 2016). The text-linguistic perspective explores the extent to which a speaker/reader will encounter each linguistic variant in texts. For example, the text-linguistic findings reported in section 4 above show that:

- all three types of genitives occur with much higher rates in written academic texts than in spoken conversation texts
- even *'s*-genitives occur with a higher rate in academic texts than in conversation texts
- the *'s*-genitive occurs with the highest rate in news texts
- but the *of*-phrase variant and premodifying noun variant occur with much higher rates in academic texts than in news texts

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

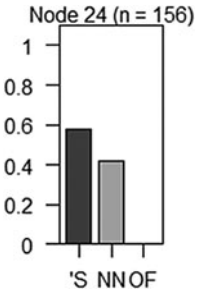
In contrast, the variationist perspective sets out to determine the factors associated with each variant, regardless of how often that variant is actually used in texts. For example, the *'s*-genitive usually occurs with animate modifying nouns, regardless of the overall textual frequency of the *'s*-variant.

Both perspectives are important, because they provide complementary information about the patterns of linguistic variation. In practice, though, it has sometimes been tempting to interpret variationist results in text-linguistic terms (or vice versa). However, as we illustrate below, this practice can lead to inaccurate conclusions. In particular, it turns out that variationist results are useful for determining the set of contextual characteristics associated with a linguistic variant – but they can be more problematic if they are interpreted in terms of the extent to which a configuration of contextual characteristics favors one variant over another.

To illustrate this point, consider the interpretation of Node 24 in the Academic CIT (figure 7). The genitive tokens in this grouping have the following contextual constraints:

Mod_semantic = Animate
Mod_sibilant = no
Mod_length = 1

This grouping has 156 total tokens; 60 percent (N = 94) of those are *'s*-genitives, and 40 percent (N = 62) are NN genitives:



It would be easy to conclude from these findings that this contextual configuration favors the choice of the *'s*-genitive in academic writing – specifically, that *'s*-genitives are more frequent than the other variants when the modifying noun is animate and the modifying NP is short. However, that conclusion would be incorrect. Rather, as we show below, NN genitives actually have a higher textual rate of occurrence than *'s*-genitives in this context.

We analyzed a total of 153 *'s*-genitives from academic writing for the variationist study (see table 1), which represented *all 's*-genitive tokens found in the academic corpus. Thus, the 94 *'s*-genitives found in this particular grouping represent the textual rate of occurrence (per 100,050 words of text) of *'s*-genitives in the academic writing corpus that occurred with short animate modifying nouns, because our annotated sample included *all* tokens of the *'s*- genitive in the academic corpus.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

In contrast, we annotated a random subsample of NN tokens from the academic corpus for the variationist study (because there were too many NN tokens to code them all). In particular, we annotated 861 of the total 1,533 NN constructions (see table 1).

As a result, the 62 NN tokens found in this CIT grouping do not represent a textual rate of occurrence. Rather, the number of observations in the CIT grouping are merely telling us how many NN tokens in our annotated subsample can be accounted for by these contextual constraints. To estimate the textual rates of occurrence (per 100,050 words of text), we can compute the proportion of NN genitives with these contextual constraints, and then extrapolate that same proportion to the entire set of NN tokens in the corpus:

Proportion of NN tokens in the sample:

62 NN tokens in this grouping / 861 annotated NN tokens = 7.2 percent
(i.e. 7.2 percent of the annotated sample of NN tokens occur with a short animate Modifying NP)

Extrapolate to a textual rate of occurrence:

.072 * 1,533 total NN tokens in the academic corpus
= 110 estimated NN genitives in the academic corpus with short animate Modifying NPs

Thus, the academic textual rate of occurrence for NN genitives with these contextual constraints (110 per 100,050 words) is actually higher than for *'s*-genitives with these contextual constraints (94 per 100,050 words).

How can we reconcile the text-linguistic findings versus variationist findings? The first takes the perspective of the text, asking which variant occurs most frequently in texts. And the second takes the perspective of the linguistic variant, describing the contextual characteristics associated with each variant. In practice, though, the predictions of the variationist perspective operate in only one direction: asking whether an occurrence of a variant predicts particular contextual characteristics. The surprising fact is that the opposite question – determining the extent to which a configuration of contextual characteristics predicts the choice among linguistic variants – is both a text-linguistic question and a variationist question.

In our example here, the variationist results show that, when the *'s*-genitive is chosen, it often occurs with a short animate modifying NP. That is, 61.4 percent of all *'s*-genitives in our sample occur with these contextual constraints (i.e. 94 of the total 153 *'s*-genitives). In contrast, when the NN genitive is chosen, it rarely occurs with a short animate modifying NP: only 7.2 percent of NN genitives occur in this context. However, the text-linguistic results show that the NN genitive with these contextual constraints actually occurs more commonly in academic texts, in part because the NN genitive is overall so much more common than the *'s*-genitive in those texts. That is, this configuration of contextual characteristics actually favors use of the NN variant over the *'s* variant in academic texts.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Results such as these indicate the need for caution in the interpretation of variationist results. Traditionally, contextual factors are treated as 'predictors' in a variationist study, with the implication that those variables can be used to predict the choice among linguistic variants. Thus, in this case study, we might conclude that the factors of an animate modifying noun coupled with a short modifying noun phrase in academic writing would 'predict' the use of the *'s*-genitive over the other genitive variants. However, the text-linguistic results show that academic writers actually choose the NN genitive more often than the *'s*-genitive in this context. Thus, the results of the variationist analysis do not really predict the choice of variant here. Those results *do* tell us what contextual factors are associated with the *'s*-genitive, when that variant is used. But, surprisingly, they do *not* tell us why an academic writer actually chooses an NN genitive more commonly than an *'s*-genitive in this context. Simply put, the variationist results show that the use of an *'s*-genitive in academic writing predicts the context of a short, animate Modifying noun phrase; but in contrast, the text-linguistic analysis shows that the context of a short, animate Modifying noun phrase does *not* accurately predict the use of an *'s*-genitive.

We are currently exploring ways to integrate these two perspectives in our ongoing research. Specifically, we are exploring ways to carry out a variationist analysis that at the same time takes account of textual rates of occurrence.

7 Discussion and conclusion

The results presented above show that the linguistic patterns of genitive variation – including both interchangeable as well as non-interchangeable (but non-categorical) tokens – can be accounted for with a high degree of accuracy. Two major contextual factors are especially important: the semantic category of the Modifying noun, and the length of the Modifying NP. In general, the *'s*-genitive is strongly associated with animate Modifying nouns, and the *of*-genitive is strongly associated with long Modifying NPs. However, beyond those two general trends, the results show a complex network of interacting factors associated with one or another of the variants, with different linguistic patterns of variation in each register.

In addition to providing a fuller linguistic description of genitive variation, the study here contributes to four larger theoretical issues:

1. The definition of the genitive construction in English and, specifically, the question of whether the NN construction constitutes a third linguistic variant functioning alongside the *'s*-genitive and postmodifying *of*-phrases
2. The question of whether *all* tokens of a linguistic variable can be accounted for (as opposed to restricting the scope of analysis to only interchangeable tokens)
3. The question of the role of register differences interacting with local contextual factors in the description of linguistic variation
4. The importance of reconciling the patterns of text-linguistic variation with the results of a variationist analysis (see section 6 above).

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Regarding the first issue, we argued (in section 1) on linguistic grounds that the NN construction constitutes a third genitive variant in English, operating as part of the same linguistic system as *'s*-genitives and postmodifying *of*-phrases. The operational definition proposed in section 1 defines a *genitive* as 'a noun phrase that functions as a restrictive modifier of a head noun phrase'. The *'s*-genitive and the NN construction fit tidily under this definition. But *of*-phrases might be regarded as modification by a prepositional phrase rather than modification by a noun phrase. The implicit assumption in our analysis here (and apparently in other studies of English genitive variation) is that the preposition *of* has no semantic content, instead serving as a grammatical function word that signals the presence of a following NP modifier.

In this regard, postmodifying *of*-phrases differ from noun modification with other prepositions in English (e.g. *in*, *on*, *under*), which often express specific semantic relations between a Head noun and a Modifying noun phrase, as in (63)–(65):

(63) a pool table in the same room
(64) the poster on the wall
(65) the water under the ice

However, several of these same prepositions have become bleached historically, so that they can also now function to simply mark the existence of a following NP modifier rather than expressing a concrete semantic relationship between the Head noun and Modifying noun (see discussion in Biber & Gray 2016: chapter 5). For example, consider (66)–(70):

(66) an increase in efficiency
(67) a significant impact on sea bird numbers
(68) Puerto Rico's status as a commonwealth
(69) the need for military unity
(70) a senior aide to the prime minister

Thus, a topic for future research is to explore these additional grammatical structures that might be regarded as genitive variants (i.e. modifying a head noun by another noun phrase).

That question relates to the second major theoretical issue that we explored in the present article: whether it is possible to account for *all* tokens of a linguistic construction, rather than restricting the analysis to only tokens judged to be interchangeable. The answer to that question is a resounding 'yes' in the case of the genitive construction.

We have argued for the importance of a third type of linguistic data that should be analyzed in variationist studies: linguistic tokens that cannot be accounted for by a categorical rule but are also not interchangeable with other variants. The statistical analyses here show that these linguistic tokens can to a large extent be accounted for on probabilistic grounds. Future research is required to explore the differences between genitive variation based on interchangeable tokens versus variation that includes all non-categorical tokens. That is, it is still not clear at this point why some non-categorical genitive tokens are judged to be interchangeable and other tokens not.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

But the present analysis shows that nearly all non-categorical genitive tokens can be accounted for in probabilistic terms.

These findings raise the possibility of extending the variationist approach to analyses of other linguistic variables defined as related structural variants that serve the same syntactic function. Possible analyses of this type include:

- the choice among grammatical devices used to postmodify a noun phrase in English (including prepositional phrases, finite relative clauses, and non-finite *-ing* relative clauses, *-ed* relative clauses, *to*-relative clauses)
- the choice among grammatical devices used as a verb complement clause (including *that*-clauses, *WH*-clauses, *to*-clauses, *-ing*-clauses; occurring in different syntactic positions)
- the choice among grammatical devices used as an adjective complement clause (including *that*-clauses, *to*-clauses, *-ing*-clauses; both extraposed and non-extraposed variants)

The standard position has been that this kind of variation is outside the scope of variationist studies, because most tokens are not strictly interchangeable with other variants. However, the results of the present study suggest that we should not exclude the possibility of such analyses on an a priori basis.

Regarding the role of register, the results of the present study strongly confirm the expectations raised by previous studies in the Text-Linguistic (TxtLx) theoretical framework (see Biber 2019; Biber & Egbert 2023), as well as the results of recent variationist studies, showing that register is 'an important factor regulating variation' (Szmrecsanyi 2019: 80). The statistical results here show that register differences matter at multiple levels, including: the set of significant predictors; the importance of predictors, both in terms of their absolute levels of importance and the relative ordering of predictors; the ways in which predictors interact; the extent to which different variants are favored by predictors; and the overall complexity of the system of constraints. Although it was not carried out as a variationist study, the results of Biber & Gray (2016) provide a detailed analysis of the historical development of these patterns, showing in particular how the NN genitive variant has evolved historically over the last three centuries, expanding both in terms of frequency of use (especially in informational written registers) and in terms of its discourse functions. Building on that body of research, we hope to further explore the underlying causes of these register differences in future research. But it is clear from the findings presented here that *register* is a hugely important predictor for any comprehensive account of linguistic variation.

https://doi.org/10.1017/S1360674323000497 Published online by Cambridge University Press

Authors' addresses:

English Department (Applied Linguistics Program)
Northern Arizona University
Flagstaff, AZ 86011-6032
USA
Douglas.Biber@nau.edu
Randi.Reppen@nau.edu
Tove.Larsson@nau.edu

Faculty of Arts (Quantitative Lexicology and Variational Linguistics)
KU Leuven
Blijde-Inkomststraat 21 – box 3308
3000 Leuven
Belgium
Benedikt.Szmrecsanyi@kuleuven.be

References

Aaron, Jessi E. 2010. Pushing the envelope: Looking beyond the variable context. *Language Variation and Change* 22, 1–36.

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

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1), 7–34.

Biber, Douglas. 2019. Text-linguistic approaches to register variation. *Register Studies* 1, 42–75.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*, 2nd edition. Cambridge: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 2023. What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies* 5, 1–22.

Biber, Douglas, Jesse Egbert, Bethany Gray, Rahel Oppliger & Benedikt Szmrecsanyi. 2016. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In Merja Kytö & Päivi Pahta (eds.), *Cambridge handbook of English historical linguistics*, 351–75. Cambridge: Cambridge University Press.

Biber, Douglas & Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 2021. *Grammar of spoken and written English*. Amsterdam: John Benjamins. [First published in 1999 by Longman.]

Brook, Marisa. 2018. Taking it up a level: Copy-raising and cascaded tiers of morphosyntactic change. *Language Variation and Change* 30, 231–60.

Egbert, Jesse & Mark Davies. 2019. If olive oil is made of olives, then what's baby oil made of? The shifting semantics of Noun+Noun sequences in American English. In Jesse Egbert & Paul Baker (eds.), *Using corpus methods to triangulate linguistic analysis*. Abingdon and New York: Routledge.

Engel, Alexandra. 2022. The register specificity of probabilistic grammars in English and Dutch. PhD dissertation, KU Leuven.
Engel, Alexandra, Jason Grafmiller, L. Rosseel & Benedikt Szmrecsanyi. 2022. Assessing the complexity of lectal competence: The register specificity of the dative alternation after *give*. *Cognitive Linguistics* 33(4), 727–66.
Engel, Alexandra & Benedikt Szmrecsanyi. 2023. Variable grammars are variable across registers: future temporal reference in English. *Language Variation and Change* 34(3), 355–78.
Fahy, Matthew, Jesse Egbert, Benedikt Szmrecsanyi & Douglas Biber. 2022. Comparing logistic regression, multinomial regression, classification trees and random forests applied to ternary variables: Three-way genitive variation in English. In Ole Schützler & Julia Schlüter (eds.), *Data and methods in corpus linguistics: Comparative approaches*, 194–223. Cambridge: Cambridge University Press.
Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3), 471–96.
Guy, Gregory. 2005. Grammar and usage: A variationist response. [Letters to *Language*.] *Language* 81(3), 561–3.
Heller, Benedikt, Benedikt Szmrecsanyi & Jason Grafmiller. 2017. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics* 45(1), 3–27.
Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11(3), 437–74.
Jankowski, Bridget & Sali Tagliamonte. 2014. On the genitive's trail: Data and method from a sociolinguistic perspective. *English Language and Linguistics* 18(2), 305–29.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4), 715–62.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania.
Labov, William. 2010. *Principles of linguistic change*, vol. 3: *Cognitive and cultural factors* (Language in Society 39). Malden, MA: Wiley-Blackwell.
Payne, John & Eva Berlage. 2014. Genitive variation: The niche role of the oblique genitive. *English Language and Linguistics* 18(2), 331–60.
Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin and New York: Mouton de Gruyter.
Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18(2), 215–62.
Szmrecsanyi, Benedikt. 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.
Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1), 76–99.
Szmrecsanyi, Benedikt & Alexandra Engel. 2023. A variationist perspective on the comparative complexity of four registers at the intersection of mode and formality. *Corpus Linguistics and Linguistic Theory* 19(1), 79–113.
Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Rothlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2), 109–37.
Tagliamonte, Sali A. 2011. *Variationist sociolinguistics: Change, observation, interpretation*. Oxford: Blackwell.
Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(02), 135–78.