Spoken syntax in a comparative perspective:

Szmrecsanyi, Benedikt; Grafmiller, Jason; Bresnan, Joan; Rosenbach, Anette; Tagliamonte, Sali; Todd, Simon

DOI:
10.5334/gjgl.310

Glossa a journal of general linguistics

Szmrecsanyi, Benedikt, et al. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2(1): 86. 1–27, DOI: https://doi.org/10.5334/gjgl.310

RESEARCH

Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English

Benedikt Szmrecsanyi[1], Jason Grafmiller[2], Joan Bresnan[3], Anette Rosenbach[4], Sali Tagliamonte[5] and Simon Todd[6]

[1] KU Leuven, Blijde-Inkomststraat 21, 3000 Leuven, BE

[2] University of Birmingham, 3 Elms Road, Edgbaston B15 2TT, UK

[3] Center for the Study of Language and Information, Stanford University, Cordura Hall 205, Stanford, CA 94305, US

[4] Tanagra Farm, P.O. Box 92, 6708 McGregor/Western Cape, ZA

[5] University of Toronto, 100 St George St, Toronto, Ontario, CDA, M5S 3G3, CA

[6] Department of Linguistics, Stanford University, Building 460, Margaret Jacks Hall, Stanford, CA 94305-2150, US

Corresponding author: Benedikt Szmrecsanyi (benszm@kuleuven.be)

This paper introduces a new resource designed to facilitate the quantitative investigation of syntactic variation in spoken language from a comparative perspective. The datasets comprise homogeneously annotated collections of "interchangeable" (i.e. competing) genitive and dative variants in four varieties of English: American English, British English, Canadian English, and New Zealand English. To showcase the empirical potential of the data source, we present a suggestive analysis that investigates the extent to which the probabilistic grammar of genitive and dative variant choice differs across varieties. The statistical analysis reveals that while there are a number of subtle probabilistic contrasts between the regional varieties under study, there is overall a striking degree of cross-varietal homogeneity. We conclude by outlining directions for future research.

Keywords: syntax; probabilistic grammar; variationist linguistics; dative alternation; genitive alternation; varieties of English

1 Introduction

In an effort to pool various pre-existing materials, work has been underway by the authors to create two comprehensive and homogeneously annotated datasets designed to facilitate the investigation of syntactic variation in spoken language from a comparative perspective.[1] Dataset 1 (henceforth: the "genitive dataset") covers the English genitive alternation, as in (1); dataset 2 (henceforth the "dative dataset") covers the English dative alternation with the verb *give*, as in (2).

(1) **The genitive alternation** (exemplification adapted from Szmrecsanyi 2006: 88)

a. [anthropology]$_{\text{possessor}}$'s [history]$_{\text{possessum}}$ is indeed implicated in the scientific construction
(the *s*-genitive)

b. it forces us to rethink […] [the history]$_{\text{possessum}}$ of [American anthropology]$_{\text{possessor}}$
(the *of*-genitive)

[1] The datasets can be freely downloaded, along with documentation, at https://purl.stanford.edu/qj187zs3852.

(2) **The dative alternation** (exemplification adapted from Tagliamonte 2014: 297)
  a. Give $[\text{me}]_{\text{recipient}}$ $[\text{some pizza}]_{\text{theme}}$
  (the ditransitive dative variant)
  b. I gave $[\text{it}]_{\text{theme}}$ to $[\text{him}]_{\text{recipient}}$
  (the prepositional dative variant)

The datasets cover four spoken regional varieties of English (or groups of spoken varieties): dialects of British English (henceforth: BrE), US American English (AmE), New Zealand English (NZE), and Canadian English (CanE). In World Englishes parlance, all of these are "central" or "inner circle" varieties of English (Kachru 1992; Mair 2013), which as advanced native varieties have reached the ultimate stage ("differentiation") in Schneider's (2007) Dynamic Model of the evolution of postcolonial Englishes.

In this paper, we introduce the new datasets to the community, and summarize the sources, the definitions of the variable contexts, annotation procedures, and so on. We use these datasets to address the extent to which the probabilistic grammars of genitive and dative choice (by which we mean the set of constraints, their effect directions and effect strengths, and the constraint rankings) differ or resemble each other across varieties of English, and we discuss how these differences or resemblances speak to theoretical questions about the usage- and experience-basedness of (knowledge of) variation. We would like to stress at the outset that the paper is not intended as the "last word", but is instead designed to inspire readers to extend our analysis in various ways.

The remainder of the paper is structured as follows. In Section 2, we provide some background on the syntactic alternations we study. Section 3 introduces the datasets. In Section 4 we investigate the overall importance of constraints on variation in a cross-variety perspective. In Section 5 we conduct careful pairwise comparisons between the varieties under study for the sake of establishing probabilistic contrasts with regard to the effect of constraints. Section 6 discusses key findings, and sketches directions for future research.

2 Background: The genitive alternation and the dative alternation

Both the genitive and the dative alternations are positional alternations: by choosing competing variants, language users can switch the order of possessor/possessum and recipient/theme. As a result, the alternations are sensitive to a number of ordering principles such as the principle of end weight (Behaghel 1909; Eitelmann 2016), given-before-new (Clifton & Frazier 2004), and – more generally speaking – Easy First (MacDonald 2013).

Historically speaking, both alternations have been available in the grammar of English for a considerable while, though the frequencies of use of the variants have changed. For the genitive alternation, we know that the *s*-genitive was used in writing much more often than the *of*-genitive up until the twelfth century (Thomas 1931: 284). This pattern reversed in the Middle English period (Mustanoja 1960: 70), to an extent that the *s*-genitive appeared to be dying out (Jucker 1993: 121). However, the *s*-genitive underwent a functional expansion in the Early Modern English period "against all odds" (Rosenbach 2002: 184), allowing both variants to survive with reasonable frequency today. For the dative alternation, the prepositional dative variant was not widely available for most of the Old English period (Mitchell 1985; Traugott 1992), but constituent order in the ditransitive construction was variable (e.g. Kemenade 1987) and sensitive to essentially the same factors that constrain the dative alternation in Modern English (De Cuypere, 2010). The prepositional dative variant began to appear in Late Old English texts (Fischer 1992), but was initially subject to lexical restrictions (Allen 2006). It developed into "a fully productive alternative" (Fischer & van der Wurff 2006: 166) to the ditransitive variant during the Middle English period, though it is still less frequent than the ditransitive today.

As to probabilistic variation in the genitive alternation across speech communities, Hinrichs and Szmrecsanyi (2007) show that in written-edited-published standard English, non-animate possessors discourage *s*-genitive usage less strongly in AmE than in BrE, and that long possessums favor *s*-genitive usage in AmE but not in BrE. Hundt and Szmrecsanyi (2012) find that possessor animacy is overall a more important predictor in (earlier, written) NZE than in (earlier, written) BrE; Jankowski and Tagliamonte (2014) report that in vernacular CanE, the *s*-genitive is near-categorically used with animate possessors, while inanimate possessors exhibit more variation and apparent-time change. Heller et al. (2017) establish that the top constraints fueling probabilistic differences across a set of nine international varieties of English include possessor animacy, constituent length, and final sibilancy.

As for the dative alternation, probabilistic differences across varieties reported in the literature include the following. Bresnan and Hay (2008) find that non-animate recipients are more likely to be used in the prepositional dative construction in spoken NZE than in spoken AmE (Bresnan & Hay 2008: Figure 1); Wolk et al. (2013) report that end-weight of themes has a stronger effect in (written) AmE than in (written) BrE; Tagliamonte (2014) diagnoses greater use of the ditransitive dative variant in BrE as well as a greater degree of spread of the ditransitive dative variant into animate recipients in contrast to CanE where the prepositional dative variant is fairly stable. Finally, with a regression model comparing each of nine varieties (including BrE, CanE and NZE) to a collective "average" variety, Röthlisberger et al. (to appear) show, among other things, that non-pronominal recipients favor the prepositional dative significantly more so in CanE than in other varieties.

3 Dataset description

To construct the datasets under study in this paper, we aimed to combine smaller, pre-existing materials (listed below) in a way that would maximize analyzability and comparability. To achieve this, we homogenized the definition of variable contexts as far as possible (excluding tokens that did not meet our guidelines for interchangeability) and standardized the annotation for the major constraints on variation (both across sub-datasets and across alternations). These efforts notwithstanding, two limitations should be spelled out at the outset. First, different sub-datasets span different ranges of real and apparent time, to the extent that this information is known. In the analysis to be presented in this paper, we take the liberty to ignore this variance, but encourage future research to include demographic information in statistical modeling, for the sake of investigating change in probabilistic grammars over time. Second, there are also some minor register differences across the sub-datasets, which we likewise ignore in the analysis to follow while acknowledging that they are potentially significant. Notice, for example, that the BrE genitive tokens were extracted from a corpus that samples oral history interviews; AmE genitives come from a corpus covering telephone conversations; and the CanE and NZE genitive tokens derive from sociolinguistic interviews. This and the more or less subtle topic differences these register differences entail is the likely reason, as we will argue in Section 5.1, why genitive *rates* (but not necessarily the underlying probabilistic grammars) vary to some extent across the sub-datasets.

With these caveats in mind, we stress that we investigated the datasets for any potential imbalances in data sampling, with a careful eye toward cross-sub-dataset asymmetries in the distributions of important predictors. In spite of the aforementioned limitations, we found a high degree of consistency among the four varieties with respect to, among other things, the distributions of possessor and possessum lengths and possessor animacy in the genitive dataset, as well as the distributions of the lengths and pronominality of recipients and themes in the dative dataset. To the extent that even small variations occurred, close inspection of the tokens themselves revealed no obvious patterns to indicate potential sampling issues. Nonetheless, we acknowledge that differences in sampling could skew our results, though the extent to which they may do so remains to be established. In what follows, we provide more information about the two datasets.

3.1 The genitive dataset

3.1.1 Sources

AmE: Most tokens were extracted from the Switchboard corpus of American English (Godfrey, Holliman & McDaniel 1992), as described in Shih et al. (2015). The Switchboard corpus covers telephone conversations collected at the beginning of the 1990s. A secondary, smaller batch of AmE genitives which is in the dataset but not included in the analysis presented in this paper, derives from the Corpus of Spoken American English (CSAE)[2] and was first analyzed in Szmrecsanyi (2006: Chapter 5). ($N = 1104$ Switchboard; $N = 319$ CSAE)

BrE: This sub-dataset derives from the Freiburg English Dialect Corpus (FRED) (Hernández 2006) and was first analyzed in Szmrecsanyi (2006: Chapter 5). It consists of oral history interviews mostly collected in the 1970s with informants that are mostly non-mobile old rural males. ($N = 1651$)

CanE: The tokens were identified in portions of sociolinguistic interviews in the Ontario Dialects Archive (ODA). The relevant materials were collected between 1997 and 2010 according to standard sociolinguistic procedures (see Jankowski & Tagliamonte 2014, though note that some additional data have been included in the present dataset). ($N = 1983$)

NZE: The tokens were extracted from sociolinguistic interviews in the Canterbury Corpus of the Origins of New Zealand English (ONZE) collection, which contains approximately 420 speakers born between 1926 and 1987, recorded between 1994 and 2008 (see Gordon, Maclagan & Hay 2007). ($N = 1953$)

The more detailed documentation accompanying the datasets provides more information about the sources. Note that the original sub-datasets were adapted to match the joint dataset's specifications as regards the definition of the variable context and annotation for constraints.

3.1.2 Definition of the variable

Only potentially alternating genitive constructions were included. In particular, only determiner *s*-genitives (i.e. *s*-genitives which have specifying function; see the documentation accompanying the dataset, where the constructions included are described meticulously)[3] and *of*-genitive constructions beginning with a definite article were considered in the analysis. For the specific selection criteria of the individual sub-datasets we refer to the references given the documentation accompanying the joint dataset. As different studies (and datasets) tend to use their own selection criteria, these sub-datasets were post-edited to make them conform to uniform criteria; for the criteria we again refer to the dataset documentation. Such post-editing could of course only eliminate constructions from the sub-datasets but not retrieve data that was excluded in the original materials, so inevitably some differences between the sub-datasets remain. The main

[2] See http://www.linguistics.ucsb.edu/research/santa-barbara-corpus.

[3] For reasons of space, we dispense with extensive exemplification in this paper, and instead refer the reader to the documentation accompanying the datasets, which features rich exemplification.

differences are documented in the dataset documentation. The NZE genitive tokens were extracted from the outset according to the guidelines mentioned above.

3.1.3 Predictors

In addition to language-external factors (variety, speaker ID, speaker sex [available for a subset of observations and not subject to analysis in the present study], speaker year of birth [available for a subset of observations and not subject to analysis in the present study]), genitive observations in the dataset were annotated for the following predictors, all of which are well-known determinants of genitive variation (see Rosenbach 2014 for discussion):

- **Response.variable:** *s*-genitive versus *of*-genitive.
- **Possessor.animacy:** The annotation distinguishes between the following categories: (1) human and animal; (2) collective; (3) temporal; (4) locative; (5) inanimate.
- **Possessor.definiteness:** (1) definite; (2) definite proper noun; (3) indefinite.
- **Semantic relation between constituents:** (1) ownership; (2) part-whole; (3) kin; (4) body part; (5) non-prototypical.
- **Possessor/Possessum.length:** length of the possessor and possessum phrases in orthographically transcribed words; determiners in the possessum phrase of *of*-genitives were subtracted from the count (as is customary in the literature), on account of the fact that a construction with a determiner *s*-genitive cannot have another determiner for structural reasons.
- **Final.sibilancy:** Presence versus absence of a final sibilant in the possessor phrase.
- **Persistence:** Specifies whether an *s*-genitive or an *of*-genitive has been used most recently within a particular window that differs by sub-dataset (see documentation for details) [not considered in the analysis to be presented in the next section].
- **Possessor/Possessum.head:** Head lexeme of both the possessor and the possessum.

The documentation accompanying the datasets provides more information about the technicalities. For the analyses to be presented in the following sections, semantic relation between constituents was modeled as a binary predictor, prototypical ((1) to (4) in the above scheme) versus non prototypical.

3.2 The dative dataset

3.2.1 Sources

AmE: The dative tokens were extracted from the Switchboard corpus of American English (Godfrey, Holliman & McDaniel 1992), as described in Bresnan et al. (2007). The Switchboard corpus covers telephone conversations collected at the beginning of the 1990s ($N = 1190$ – we reiterate that here and in the other sources attention is restricted to the dative verb *give*).

BrE and CanE: Datives were extracted from materials collected in the UK and Canada between 1997 and 2010 according to standard sociolinguistic procedures. In the UK the data come from York and small towns and villages all over the UK. In Canada, the data come from Toronto, as well as numerous small towns and villages in Ontario (see Tagliamonte 2014). We add that the Canadian data has an apparent time difference (BrE: $N = 944$; CanE: $N = 1157$).

NZE: The observations come from the Origins of New Zealand English corpora (ONZE) (see Gordon, Maclagan & Hay 2007), and were collected separately but overlap substantially with those presented by Bresnan and Hay (2008). Note that the New Zealand dative sub-dataset includes historical data from early New Zealand English speakers, and so covers speakers born between 1851 and 1984 ($N = 845$).

The documentation accompanying the datasets provides more information about the sources. Note that the original sub-datasets were adapted to match the joint dataset's specifications as regards the definition of the variable and annotation for constraints.

3.2.2 Definition of the variable

Attention was restricted to the dative verb *give*. The definition of interchangeable ditransitive and prepositional dative variants broadly follows Bresnan et al. (2007), which essentially means that all instances of *give* with two argument NPs minus non-interchangeable constructions were considered. The documentation accompanying the dataset provides more information.

3.2.3 Predictors

In addition to language-external factors (variety, speaker ID, speaker sex [available for a subset of observations and not subject to analysis in the present study], speaker year of birth [available for a subset of observations and not subject to analysis in the present study]), dative observations in the dataset were annotated for the following predictors, which are well-known determinants of dative variation (see Bresnan & Ford 2010 for discussion):

- **Response.variable:** Ditransitive dative versus prepositional dative. (In addition, the dataset also documents a small number ($N = 8$) of non-standard patterns of the type *he had to give it her*, which were excluded in the analysis to be presented in this paper).
- **Recipient/Theme.type:** The annotation distinguishes between the following categories: (1) noun phrase; (2) personal pronoun; (3) demonstrative pronoun; (4) impersonal pronoun.
- **Recipient/Theme.definiteness:** The annotation distinguishes between the following categories: (1) definite; (2) indefinite (3) definite proper noun.
- **Recipient/Theme.animacy:** The annotation distinguishes between the following categories: (1) human and animal; (2) collective; (3) temporal; (4) locative; (5) inanimate.
- **Recipient/Theme.length:** Length of the recipient and theme phrases in orthographically transcribed words.
- **Semantics (of dative verb):** (1) transfer; (2) communication; (3) abstract.
- **Recipient/Theme.head:** Head lexeme of both the theme and the recipient.

The documentation accompanying the datasets provides more information about the technicalities. For the analyses to be presented in the following sections, the predictors were simplified as follows: Recipient/Theme.type were reduced to a binary contrast: pronominal ([2], [3], [4]) versus non-pronominal ([1]); Recipient/Theme.definiteness were reduced to a binary contrast: definite ([1], [3]) versus indefinite ([2]); Recipient/Theme.animacy were reduced to a binary contrast: animate ([1]) versus inanimate ([2], [3], [4], [5]); and the Recipient/Theme.length measures were combined into a relative measure of length (henceforth: Length.difference), calculated as *log*(Recipient.length) - *log*(Theme.length) following Bresnan and Ford (2010).

4 The bird's eye perspective: Establishing constraint importance across varieties

To set the stage, this section ranks individual constraints per variety and alternation according to explanatory importance, using Conditional Random Forest (CRF) analysis (see Tagliamonte & Baayen 2012 for an accessible introduction). A desirable property of CRF models is that they are robust to problems of predictor correlation, as their predictions are derived from ensembles of conditional inference trees grown on randomly sampled subsets of the data (Breiman 2001; Strobl, Malley & Tutz 2009). Each tree is built on a random subset of predictors whose explanatory power is assessed through random permutation of their values, which breaks the predictor's association with the outcome. A "permutation variable importance" measure for each predictor can then be calculated by measuring the decrease in accuracy of the model with the permuted predictor compared to the accuracy of the model with the non-permuted predictor, and used to rank predictors. We conducted an independent CRF run for each of the varieties, per alternation, utilizing the cforest() function in R's party package (Hothorn, Hornik & Zeileis 2006; Strobl et al. 2008).

The variable importance measures for the genitive alternation are shown in Figure 1. Possessor animacy is the most important predictor in all varieties by a wide margin; possessor length is the next most important predictor in all varieties except in CanE (Figure 1c), a variety in which possessor animacy is supremely important but other constraints are largely not. According to CRF, other constraints have only a minor influence on genitive choice overall, and their relative ranking tends to vary across varieties. For example, it appears that final sibilancy is relatively important in NZE (Figure 1d), but much less so in AmE (Figure 1a) and BrE (Figure 1b).

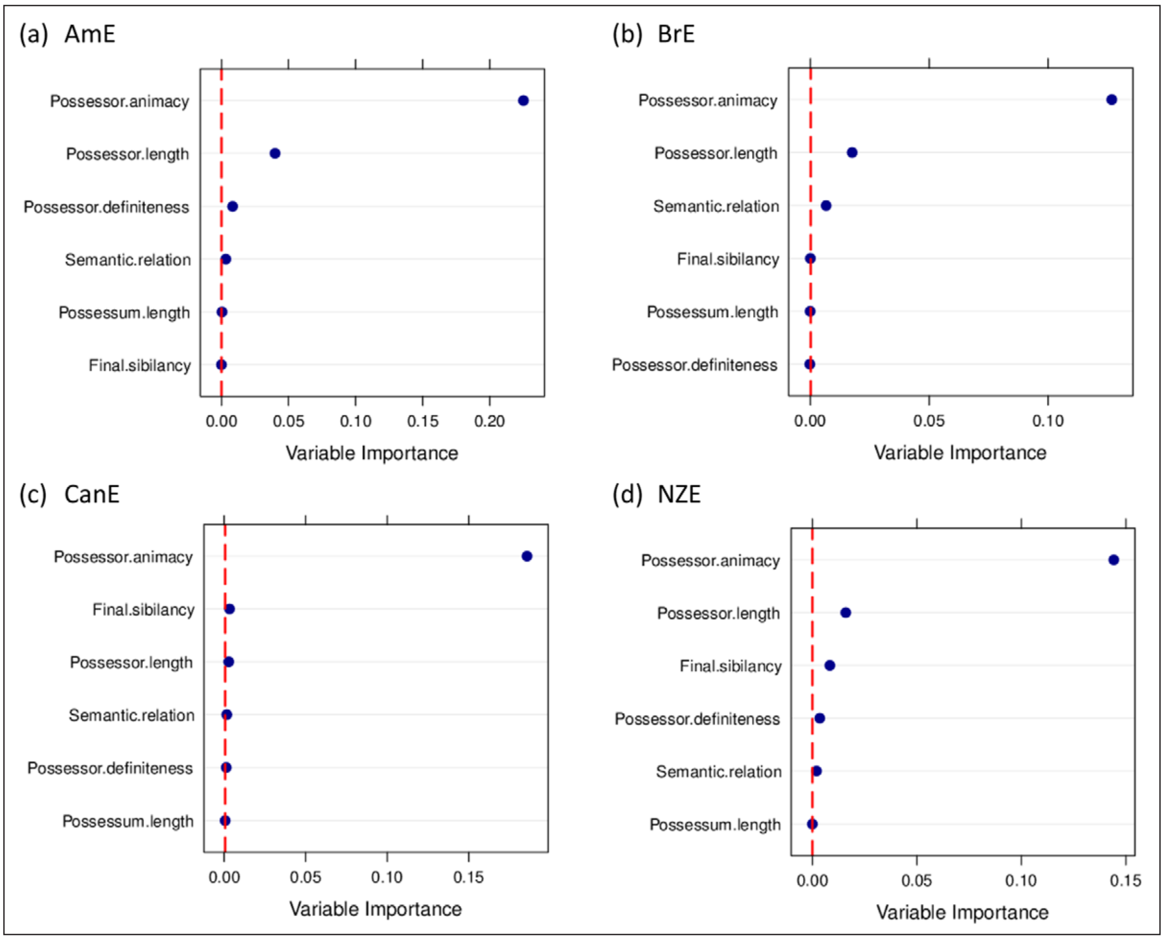

Figure 1: CRF permutation variable importance measures for the four genitive sub-datasets.

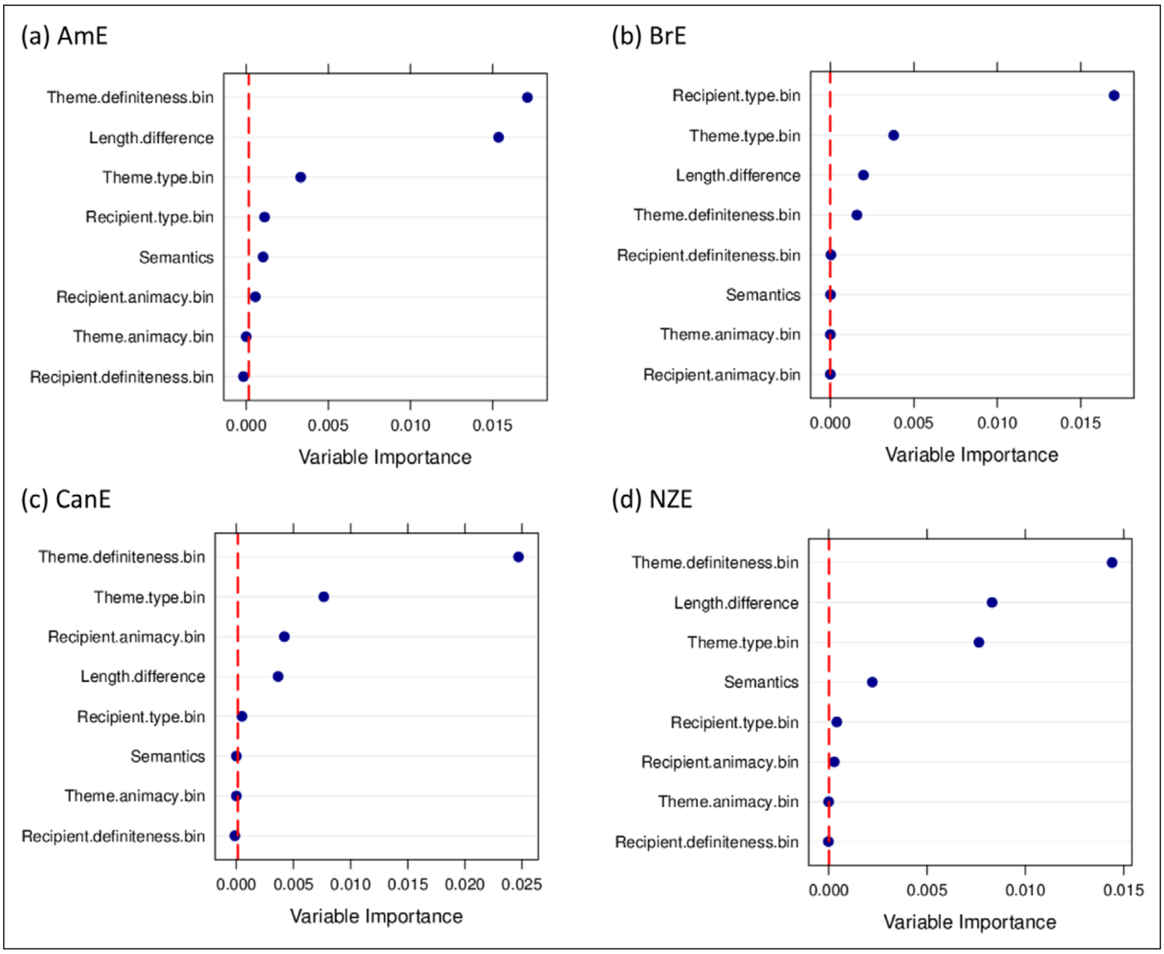

Figure 2: CRF permutation variable importance measures for the four dative sub-datasets.

The corresponding variable importance rankings for the dative alternation are shown in Figure 2. Compared to the genitive alternations rankings, more constraints are crucially implicated in variation patterns; there also seems to be more variability between varieties. That said, recipient/theme type and theme definiteness are consistently among the top-ranked predictors, while recipient definiteness is rather unimportant, except to some degree in BrE. AmE (Figure 2a) is different from the other varieties thanks to the important role that the length difference between recipient and theme plays. CanE (Figure 2c) is special because recipient animacy is comparatively important.

In summary, we would like to highlight the comparative similarity of the genitive alternation rankings and the relative dissimilarity of the dative alternation rankings: in the genitive alternation, possessor animacy is the uncontested prime constraint (which is of course not surprising given the literature – see e.g. Rosenbach 2005) wherever we look, while the situation is more "pluralistic", as it were, in the dative alternation. However, some stability also exists across varieties for the dative alternation, in that recipient/theme type and theme definiteness are consistently among the top-ranked constraints.

5 The jeweler's eye perspective: Pairwise regression modeling

In this section we take a close look at differences between the varieties under study regarding the effect size and direction of the various constraints on genitive and dative choice. The analysis technique we use is binary logistic regression analysis with mixed effects (Gelman & Hill 2007; Zuur et al. 2009) as implemented in the lme4 package in R (R Core Team 2014; Bates et al. 2015). The investigation proceeded in a round-robin fashion, where the data from each variety were compared to the data from each other variety, resulting in a total of six regression runs per alternation comprising every possible pairwise combination. Compared to an alternative unitary model pooling all varieties for each alternation and incorporating interaction terms between variety and the language-internal constraints, our approach adds considerable resolution because it is more sensitive to pairwise probabilistic differences between varieties than the unitary-model approach, in which one would have to define a default variety or use other contrast designs.

Fixed effect predictors in regression analysis are those mentioned in the preceding sections. All models also initially contained an interaction of the variable "variety" (BrE vs. AmE vs. NZE vs. CanE) with each of the language-internal predictors.[4] No higher order interactions were included. The initial random effects structure, prior to pruning, consisted of random intercepts for speaker (idiolectal differences) as well as random intercepts for the head nouns of the possessor and possessum/recipient and theme NPs (note that following Wolk et al. 2013: 399, we collapsed all speaker and head noun levels that did not reach a threshold of at least five observations). The same initial model structure was fit to each of the six pairwise sub-datasets, and was then individually optimized following the top-down procedure outlined by Zuur et al. (2009: 120–122). Starting with the maximal model structure, the contribution of each random effect was first assessed via likelihood ratio tests. After settling on the optimal random structure, we removed non-significant interaction terms whose standard errors were larger than the absolute value of their coefficients.[5] Standard model diagnostics were then applied to each model, checking for overdispersion, leverage, residual structure and normality of random effects (Harrell 2001; Gelman & Hill 2007; Baayen 2008). Lastly, models were run through a 10-fold cross-validation procedure to check for the possibility of overfitting. Multicollinearity is not a serious issue, as the condition number κ in all pairwise regression models is below the threshold of 15 which indicates medium collinearity (Baayen 2008: 200).

In what follows we summarize pairwise comparisons that exhibit significant probabilistic contrasts. Detailed model outputs are available as supplementary materials (supplementary file 1).

5.1 The genitive alternation

5.1.1 Comparing BrE to CanE

We begin with a comparison of constraints on genitive choice among speakers of CanE and BrE. In these varieties, we find approximately equivalent proportions of genitive variants across the two varieties (Table 1). The CanE/BrE regression model (in which CanE serves as the default variety) fits the data very well ($C = 0.98$), and correctly predicts 93.2% of the observations, from a baseline of 63.1%. All language-internal constraints have main effects in the directions predicted by the literature (see e.g. Rosenbach 2014): animate possessors favor the *s*-genitive, as do definite possessors, and longer possessums; the pres-

Table 1: Genitive variant rates in the CanE and BrE sub-datasets.

	***of*-genitive**	**%**	***s*-genitive**	**%**	**Total**
CanE	706	35.6	1277	64.4	1983
BrE	636	38.5	1015	61.5	1651
Total	1342		2292		3634

[4] Preliminary data exploration of the genitives dataset suggested possible interactions of animacy and constituent length, so interactions of animacy with both possessor and possessum length were also included in the initial models.

[5] This is a more conservative criterion for elimination than the usual $p < .05$, as it only removes predictors whose estimated effects are extremely small (or variance in the estimate is extremely large) and therefore unlikely to be affected by the presence/absence of other factors in the model.

ence of a final sibilant disfavors the *s*-genitive, while a prototypical relation between possessor and possessum (e.g. kinship or ownership) favors it; and so on. This pattern is consistent across all the pairwise genitive models, so we will not discuss main effects hereafter.

As for probabilistic contrasts, the model reveals a probabilistic contrast between CanE and BrE with regard to the effect of possessor animacy: BrE speakers' genitive construction choices are less sensitive to possessor animacy (as measured by the probabilistic difference which an animate possessor makes, compared to an inanimate possessor) than those of CanE speakers.

5.1.2 Comparing BrE to NZE

We now turn to a comparison of BrE and NZE, where, unlike in the previous case, we do find substantial cross-varietal differences in the overall frequencies of genitive variants (Table 2). According to our data, NZE speakers use the *s*-genitive notably less often than speakers of BrE (see this section's Interim summary for a discussion of this frequency differences).

The BrE/NZE model (default variety: NZE) fits the data very well ($C = 0.986$), correctly predicting 94.1% of the observations, from a baseline of 52.8%. Two language-internal constraints interact significantly with variety. First, as in the CanE/BrE comparison, we see a difference between BrE and NZE with regard to the effect of possessor animacy, which has a stronger effect in NZE than in BrE: what we are seeing is that BrE speakers' genitive construction choices are less sensitive to the difference between inanimate and animate possessors than those of NZE speakers.

Second, the effect of a final sibilant in the possessor is significantly stronger in NZE than in BrE, in that NZE speakers avoid the *s*-genitive in the presence of a final sibilant to a larger extent than do BrE speakers. The partial effects plot in Figure 3 visually depicts the contrast; note how the NZE line is steeper than the BrE line.

5.1.3 Comparing CanE to NZE

Let us set aside BrE for the time being, and turn to the third comparison for which we find cross-varietal differences: CanE versus NZE. In terms of overall frequency, here too we find varying proportions of genitive variants across the two sub-datasets (Table 3): CanE speakers use the s-genitive considerably more often than speakers of NZE do.

The fit of the regression model (default variety: CanE) is very good ($C = 0.989$; accuracy = 94.8%, baseline = 50.2%). There are two significant interactions between variety and language-internal constraints on genitive choice. First, similar to the BrE/NZE comparison the effect of a final sibilant in the possessor is significantly stronger in NZE than in CanE in that NZE speakers avoid the *s*-genitive in the presence of a final sibilant to a larger extent than CanE speakers do. Second, the effect of the semantic relations between the genitive constituents is significantly stronger in NZE than in CanE: when choosing genitive variants, NZE speakers are more sensitive to the difference between prototypical and non-prototypical semantic relations than CanE speakers are. The difference is visually depicted in the partial effects plot in Figure 4.

Table 2: Genitive variant rates in the BrE and NZE sub-datasets.

	***of*-genitive**	**%**	***s*-genitive**	**%**	**Total**
BrE	636	38.5	1015	61.5	1651
NZE	1268	64.9	685	35.1	1953
Total	1904		1700		3604

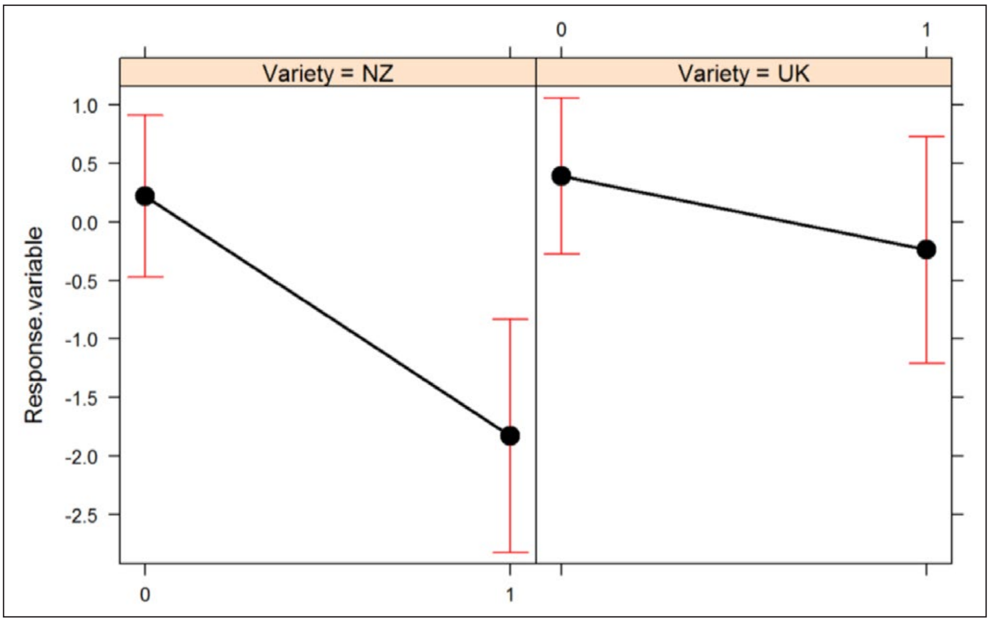

Figure 3: Partial effects plots of the interaction between variety and final sibilancy (“0” – absent, “1” – present) in the BrE/NZE model. Log odds (Response.variable) are for the *s*-genitive.

Table 3: Genitive variant rates in the CanE and NZE sub-datasets.

	***of*-genitive**	**%**	***s*-genitive**	**%**	**Total**
CanE	706	35.6	1277	64.4	1983
NZE	1268	64.9	685	35.1	1953
Total	1974		1962		3936

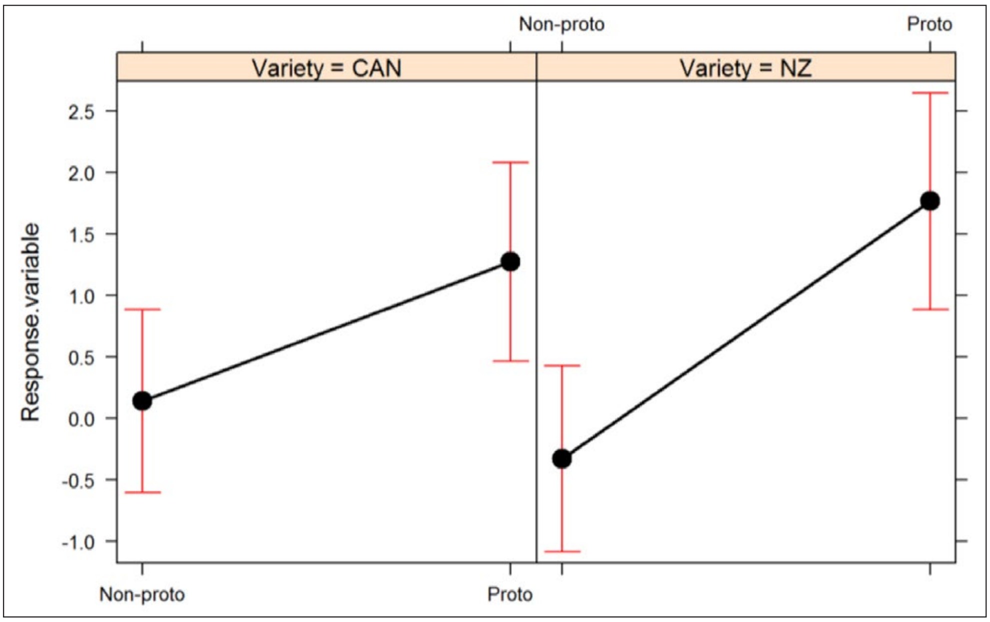

Figure 4: Partial effects plots of the interaction of variety with semantic relation (“Non-proto” – non-prototypical; “Proto” – prototypical) in the CanE/NZE model. Log odds (Response.variable) are for the *s*-genitive.

5.1.4 Remaining comparisons

Beyond the aforementioned interaction effects, we find no other significant differences between varieties in the three remaining comparisons (AmE/BrE, AmE/CanE, AmE/NZE). We thus take the liberty to skip a discussion of these models, and refer the reader to the supplementary materials instead.

5.1.5 Genitive variation: Interim summary

We begin by discussing why we see differences in genitive rates: recall that the *s*-genitive is strikingly more frequent in CanE (64.4%) and BrE (61.5%) than in AmE (40.5%) and NZE (35.1%). It seems that these differences primarily boil down to topic differences: as we have seen in Section 4, the most important constraint on genitive variation by far is possessor animacy. Animate possessors favor the s-genitive, and it so happens that in the sociolinguistic interviews and oral history materials that make up our CanE and BrE sub-datasets, people talk more about animate NPs than in the AmE and NZE sub-datasets: in the CanE sub-dataset, a full 66.8% of all possessors are animate, and in the BrE sub-dataset, the figure 65.2%. By contrast, the share of animate possessors is only 37.6% and 40.7%, respectively, in the NZE and AmE sub-datasets. These frequency differences are certainly interesting, but we stress that they do not necessarily speak to the issue of probabilistic variation patterns. Instead, we are dealing here with what Szmrecsanyi (2016) calls an "environmental" difference: the exact same probabilistic grammar will yield different variant rates if the input frequency of relevant stimuli (e.g. animate possessors) is sufficiently dissimilar. But such differences do not tell us much about the extent to which language users may or may not have different ways of choosing between variants.

Turning to actual probabilistic differences, Table 4 summarizes the four significant probabilistic contrasts that have surfaced in the foregoing analysis. Animacy of the possessor is implicated in two of the five contrasts, and so we may conclude that possessor animacy is not only one of the most important constraints on genitive variation, but also one of the most variable (see also Rosenbach 2017).[6] Final sibilancy likewise takes part in two comparisons. More generally speaking, we see that the probabilistic genitive grammars under study exhibit subtle yet measurable differences not only with regard to the ranking of the constraints (as determined in CRF analysis – see Figure 1) but also in terms of effect size.

Looking at the varieties that are involved in probabilistic contrasts, observe that NZE is implicated in four of the five contrasts we diagnosed. One pattern, in this connection, is that final sibilancy tends to have a stronger effect in NZE than in some other varieties. The pattern that emerges with regard to BrE is that the influence of possessor animacy appears to be notably weaker in BrE than in other varieties.

What is the extent to which the contrasts uncovered in the pairwise analyses are consonant with the previous literature? To begin with, our finding that the effect of possessor animacy is stronger in NZE than in BrE ties in with Hundt and Szmrecsanyi (2012), however their significant effects were limited to collective and locative, rather than animate,

[6] The interpretation of interaction terms involving possessor animacy is to some extent contingent on the fact that the default level for animacy was set to "inanimate" in regression analysis, hence the models quantify the difference that e.g. animate possessors make vis-à-vis inanimate possessors. That said, the distribution of genitives with non-animate possessors indicates differences that are not well captured in our regression modeling: in particular, the NZE data stand out in showing a great reluctance to use the *s*-genitive with non-animate possessors (collapsing collective, locative, temporal and other inanimate possessors) compared to the other varieties, as reflected in the relative frequencies of the *s*-genitive with non-animate possessors (NZE: 4.0%; CanE: 8.7%; BrE: 10.8%; AmE: 10.7%). A more in-depth study of this issue is reserved for another occasion.

Table 4: Significant probabilistic contrasts in pairwise regression modeling of the genitive alternation.

variety pairing	constraint	*p*	description
BrE/CanE	Possessor.animacy	$p < 0.05$	Effect of possessor animacy is stronger in CanE than in BrE: CanE speakers are more sensitive to the difference between inanimate and animate possessors than BrE speakers are.
BrE/NZE	Possessor.animacy	$p < 0.05$	Effect of possessor animacy is stronger in NZE than in BrE: NZE speakers are more sensitive to the difference between inanimate and animate possessors than BrE speakers are.
	Final.sibilancy	$p < 0.01$	Effect of final sibilancy is stronger in NZE than in BrE: NZE speakers avoid the *s*-genitive in the presence of a final sibilant more than BrE speakers do.
CanE/NZE	Final.sibilancy	$p < 0.05$	Effect of final sibilancy is stronger in NZE than in CanE: NZE speakers avoid the *s*-genitive in the presence of a final sibilant more than CanE speakers do.
	Semantic relation between constituents	$p < 0.05$	Effect of Semantic relation is stronger in NZE than in CanE: NZE speakers favor the *s*-genitive more strongly with prototypical relations than CanE speakers do.

possessors (again, vis-a-vis inanimate possessors). We will not speculate as to why our results are different from Hundt and Szmrecsanyi (2012) except to note that their study examined data of a different time period (late 19th and early 20th century) and medium (written) from ours. Next, we saw earlier that previous studies comparing AmE to BrE have reported a number of probabilistic contrasts, such as inanimate possessors discouraging *s*-genitive usage less strongly in AmE than in BrE, and/or animate possessor more strongly favoring the *s*-genitive in BrE than in AmE (Jahr Sohrheim 1980; Hinrichs & Szmrecsanyi 2007; Szmrecsanyi & Hinrichs 2008; Szmrecsanyi et al. 2014). The previous literature also suggests that in the written medium at least, long possessums favor *s*-genitive usage in AmE but not in BrE (Hinrichs & Szmrecsanyi 2007). However, our pairwise regression-based comparison between BrE and AmE fails to detect any significant probabilistic differences with regard to genitive choice. The most likely reason for this failure is that the present study examines spontaneous spoken data, while Hinrichs & Szmrecsanyi (2007) examined written-edited-published language from sections of the Brown corpora. Recent work has found sizeable differences in genitive usage across spoken and written modes and registers, even within the same variety (Grafmiller 2014), thus the results presented here should be interpreted within the larger context of what we know about genitive variation across different modes, registers and communities.[7] Finally, we note that none of the probabilistic contrasts that our analysis unearths emerge as significant in Heller et al. (2017), a recent study on probabilistic genitive variation that covers the same varieties that are represented in Table 4, albeit using different multivariate designs and on the basis of different, more acrolectal and less vernacular data sources (the International Corpus of English and the Corpus of Global Web-Based English).[8]

[7] As an aside, we note that in the BrE data a good number of interviewees are fishermen, who talk a lot about their fishing boats. Since Early Modern English, boats and ships have had a tendency to be encoded with the *s*-genitive (Breejen 1937: 55). Future study may want to investigate the extent to which this peculiarity of the BrE dataset may conceal actual probabilistic differences between BrE and AmE.

[8] The only effect that seems to shine through, upon closer inspection, in the spoken-only subpart of the Heller et al. dataset is the final sibilancy difference between BrE and NZE, which happens to be the only contrast in Table 4 that is significant at the $p < 0.01$ level.

Table 5: Dative variant rates in the CanE and NZE sub-datasets.

	prepositional dative	%	ditransitive dative	%	Total
BrE	82	8.7	862	91.3	944
AmE	160	13.4	1030	86.6	1190
Total	242		1892		2134

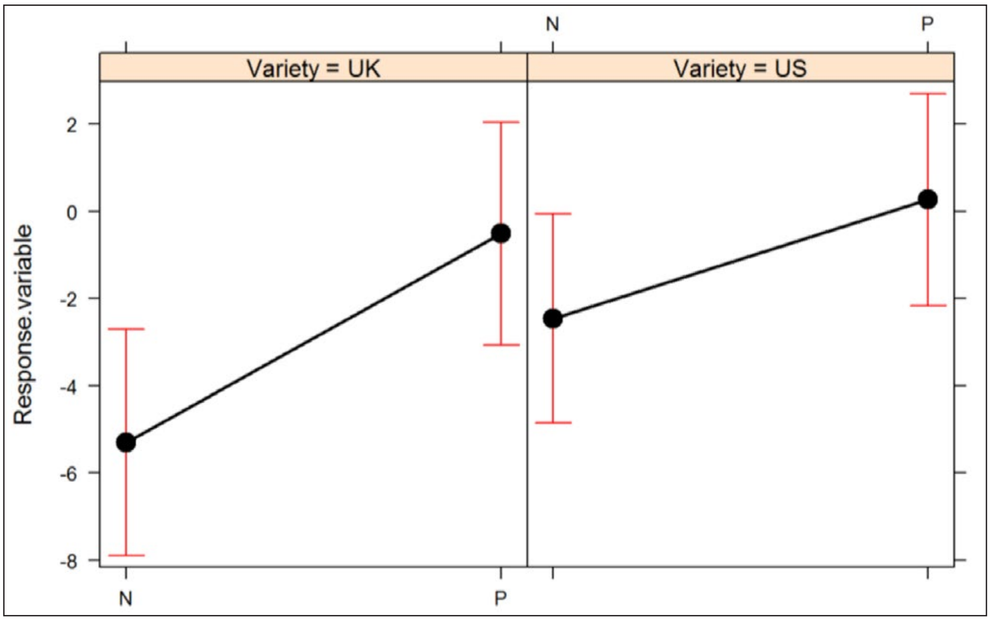

Figure 5: Partial effects plot of the interaction of variety with recipient type ("N" – nominal; "P" – pronominal) in the BrE/AmE model. Log odds (Response.variable) are for the prepositional dative.

5.2 *The dative alternation*

5.2.1 Comparing AmE to BrE

A frequency analysis reveals that the ditransitive dative variant is overall a bit more popular in BrE than in AmE (Table 5). The pairwise regression model fit on the combined sub-dataset (default variety: BrE) has a very good fit ($C = 0.987$) and correctly predicts 96.4% (baseline: 88.8%) of all dative outcomes. The constraints all have the effect directions predicted by the literature: longer recipients favor the prepositional dative variant, as do non-pronominal recipients, pronominal themes, indefinite recipients, and definite themes – and so on. This pattern is fairly consistent across all the pairwise dative models, so we will not discuss main effects in dative variation hereafter.

Regression analysis diagnoses one significant probabilistic contrast between BrE and AmE: in both varieties, non-pronominal recipients favor the prepositional dative variant (vis-à-vis pronominal recipients), but the effect is weaker in AmE than in BrE. The difference is shown in Figure 5 (notice the steeper line in the UK diagram).

5.2.2 Comparing AmE to CanE

Variant proportions are similar in CanE and AmE (Table 6). The pairwise regression model (default variety: CanE) has a very good fit ($C = 0.980$) and correctly predicts 96.1% (baseline: 85.3%) of all dative outcomes.

We find two significant interaction effects. First, in CanE communication uses of *give* disfavor the prepositional dative compared to transfer verbs whereas in AmE communication tokens actually favor the prepositional dative, vis-a-vis transfer tokens (this difference is visually depicted in Figure 6). Second, in both varieties longer recipients favor the prepositional dative variant – as expected per the principle of end weight – but the effect is stronger in AmE than in CanE. In other words, the principle of end weight is a more powerful determinant of variant choice in AmE than in CanE. Figure 7 illustrates the difference.

Table 6: Dative variant rates in the CanE and NZE sub-datasets.

	prepositional dative	%	ditransitive dative	%	Total
CanE	185	16.0	972	84	1157
AmE	160	13.4	1030	86.6	1190
Total	345		2002		2347

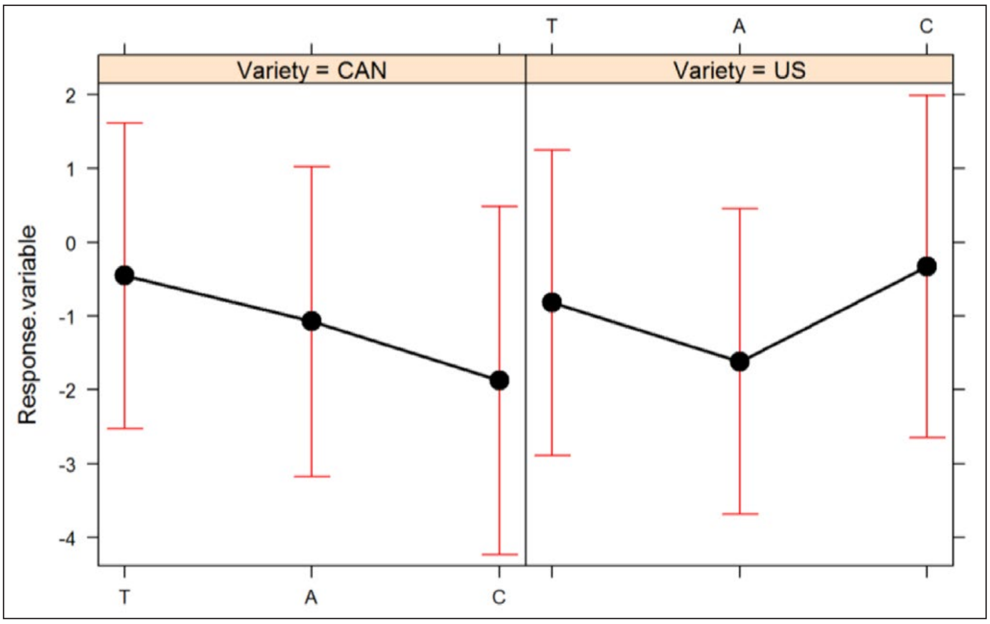

Figure 6: Partial effects plot of the interaction of variety with semantics of *give* ("T" – transfer; "A" – abstract; "C" – communication) in the AmE/CanE model. Log odds (Response.variable) are for the prepositional dative.

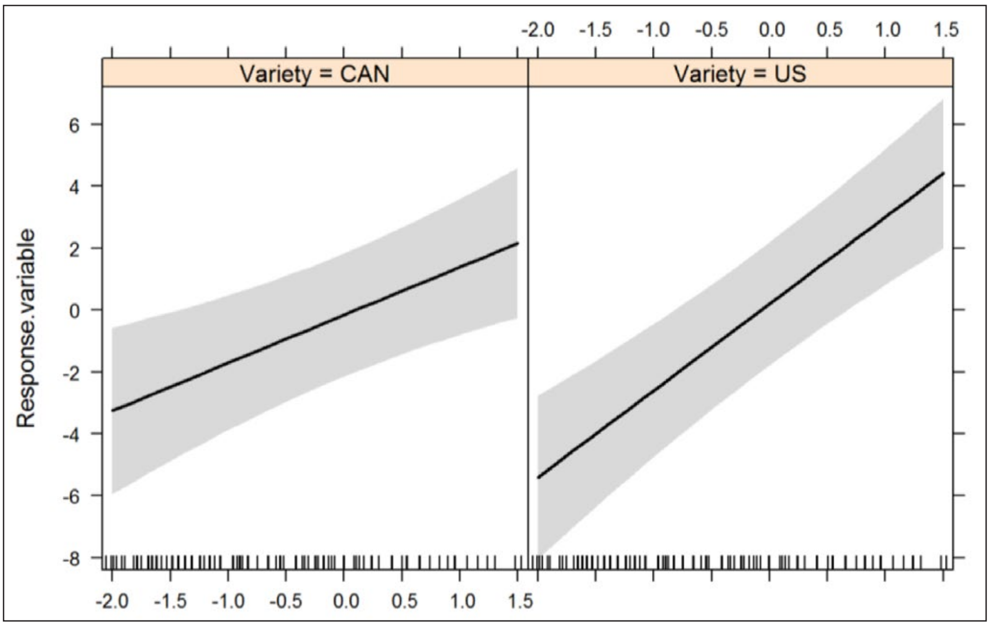

Figure 7: Partial effects plot of the interaction of variety with the Length.difference between recipient and theme in the AmE/CanE model. Log odds (Response.variable) are for the prepositional dative.

5.2.3 Comparing AmE to NZE

Table 7 indicates that there are no major frequency differentials. The regression model (default variety: NZE) based on the combined NZE/AmE sub-dataset has a very good fit ($C = 0.990$) and correctly predicts 97.0% of all dative outcomes (baseline: 87.9%).

The regression model identifies a significant interaction between variety and the semantics of *give*: in NZE, communication uses of *give* disfavor the prepositional dative compared to transfer verbs, but in AmE communication tokens actually marginally favor the prepositional dative, compared to transfer tokens. This is parallel to the pattern that we saw in the AmE/CanE comparison (see Figure 6).

5.2.4 Comparing BrE to CanE

Comparing variant rates in CanE and BrE, we find that the prepositional dative variant is somewhat more popular in CanE than in BrE (see Table 8). The regression model based on the CanE/BrE sub-dataset (default variety: NZE) has a good fit ($C = 0.979$) and correctly predicts 96.1% of all outcomes, against a baseline of 87.4%.

We find two significant interaction effects between variety and language-internal constraints. First, as recipient length increases relative to theme length, the prepositional dative variant is favored more strongly in BrE than in CanE. In other words, the principle of end weight is more powerful in BrE than in CanE. This is similar to the patterns we saw in the AmE/CanE comparison (see Figure 7). Second, vis-à-vis pronominal recipients nominal recipients favor the prepositional dative variant more robustly in BrE than in CanE. Although this effect is naturally correlated with that of recipient length (pronominal recipients always being very short), tests for data multicollinearity provide reasonable evidence that the two effects are independent.

5.2.5 Comparing BrE to NZE

In NZE, variant proportions are roughly similar to those in BrE (Table 9). The regression model (default variety: NZE) has a very good fit ($C = 0.992$) and correctly predicts 97.3% of all dative outcomes (baseline: 90.7%).

Table 7: Dative variant rates in the NZE and AmE sub-datasets.

	prepositional dative	%	ditransitive dative	%	Total
NZE	89	10.5	756	89.5	845
AmE	160	13.4	1030	86.6	1190
Total	249		1786		2035

Table 8: Dative variant rates in the CanE and BrE sub-datasets.

	prepositional dative	%	ditransitive dative	%	Total
CanE	185	16.0	972	84.0	1157
BrE	82	8.7	862	91.3	944
Total	267		1834		2101

Table 9: Dative variant rates in the BrE and NZE sub-datasets.

	prepositional dative	%	ditransitive dative	%	Total
BrE	82	8.7	862	91.3	944
NZE	89	10.5	756	89.5	845
Total	171		1618		1789

There is one significant probabilistic contrast: in NZE, communication uses of *give* disfavor the prepositional dative compared to transfer uses (the default category); there is also a tendency in NZE that abstract uses disfavor the prepositional dative. In BrE, by contrast, abstract uses favor the prepositional dative (again, compared to transfer uses), while communication uses tend to have no effect. Figure 8 visually depicts the interaction.

5.2.6 Comparing CanE to NZE

As can be seen from Table 10, the prepositional dative is moderately more frequent in CanE than in NZE. The regression model based on the combined sub-datasets (default variety: CanE) has a very good fit ($C = 0.980$) and correctly predicts 96.8% (baseline: 86.4%) of all dative outcomes.

The model diagnoses one significant interaction effect: the effect of length differences between recipient and theme is different in the two varieties. In both varieties, longer recipients favor the prepositional dative variant – as expected per the principle of end weight – but the effect is stronger in NZE than in CanE (similar to what we saw in the AmE/CanE comparison – see Figure 7). In other words, the principle of end weight is a more powerful determinant of variant choice in NZE than in CanE.

5.2.7 Dative variation: Interim summary

Table 11 summarizes the eight significant probabilistic contrasts identified in pairwise regression modeling. Adopting a constraint-centered perspective, our findings concern-

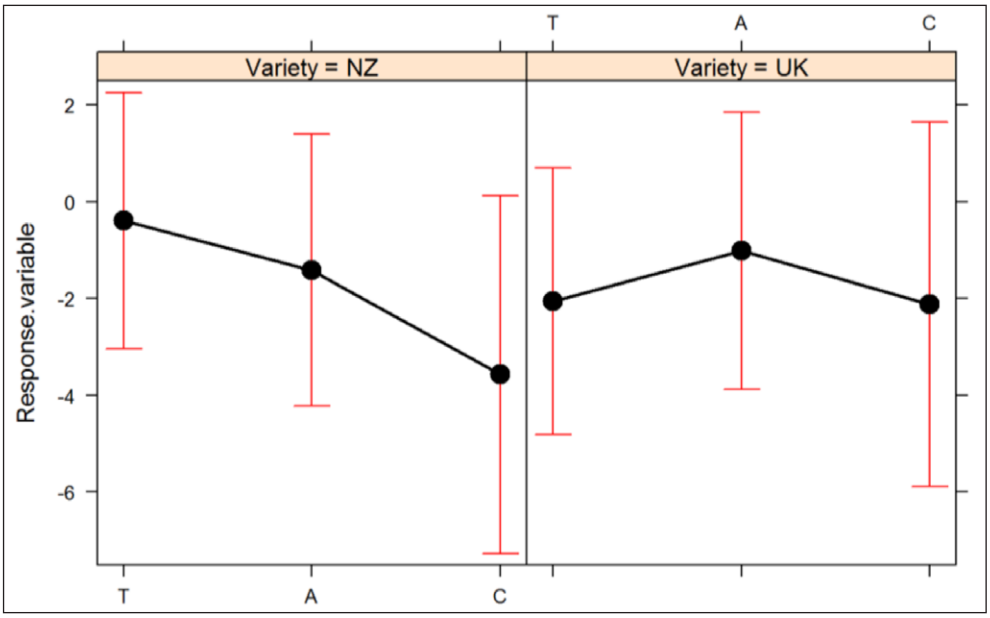

Figure 8: Partial effects plot of the interaction of variety with the semantics of *give* ("T" – transfer; "A" – abstract; "C" – communication) in the AmE/CanE model. Log odds (Response.variable) are for the prepositional dative.

Table 10: Dative variant rates in the CanE and NZE sub-datasets.

	prepositional dative	%	ditransitive dative	%	Total
CanE	185	16.0	972	84,0	1157
NZE	89	10.5	756	89.5	845
Total	274		1728		2002

Table 11: Significant probabilistic contrasts in pairwise regression modeling of the dative alternation.

variety pairing	constraint	*p*	description
AmE/BrE	Recipient.type	$p < 0.005$	Non-pronominal recipients favor the prepositional dative variant less strongly in AmE than in BrE.
AmE/CanE	Semantics of the dative verb	$p < 0.05$	In CanE, communication uses of *give* disfavor the prepositional dative compared to transfer tokens, but in AmE communication tokens actually favor the prepositional dative, compared to transfer tokens.
	Length.difference	$p < 0.05$	In both varieties, longer recipients favor the prepositional dative variant (as predicted by the principle of end weight) but the effect is stronger in AmE than in CanE.
AmE/NZE	Semantics of the dative verb	$p < 0.005$	In NZE, communication uses of *give* disfavor the prepositional dative compared to transfer tokens, but in AmE communication tokens actually favor the prepositional dative, compared to transfer tokens.
BrE/CanE	Recipient.type	$p < 0.001$	Non-pronominal recipients favor the prepositional dative variant less strongly in CanE than in BrE.
	Length.difference	$p < 0.05$	In both varieties, longer recipients favor the prepositional dative variant (as predicted by the principle of end weight) but the effect is stronger in BrE than in CanE.
BrE/NZE	Semantics of the dative verb	$p < 0.05$	In BrE abstract uses of *give* favor the prepositional dative compared to transfer uses (the default category); but in NZE, abstract uses *dis*favor the prepositional dative compared to transfer tokens.
CanE/NZE	Length.difference	$p < 0.005$	In both varieties, longer recipients favor the prepositional dative variant (as predicted by the principle of end weight) but the effect is stronger in NZE than in CanE.

ing the dative alternation may be summarized as follows. The semantics of the dative verb make a significant difference in three comparisons, and what we see is in fact different effect directions across varieties – for example, in AmE communication tokens of *give* tend to favor the prepositional dative, unlike in other varieties. It might be expected that one would see probabilistic differences like this particularly with constraints such as verb semantics, which are perhaps particularly amenable to cultural/regional (re-) interpretation. This set of contrasts clearly merits more scrutiny, and future research is encouraged to investigate the issue further. Length effects are likewise involved in three contrasts; here the very consistent pattern is that end weight effects are weaker in Canadian English than in all other varieties. Lastly, Recipient.type surfaces in two comparisons, and in each case the prepositional dative-favoring effect of non-pronominal recipients is stronger in BrE than in other varieties.

What is the extent to which these patterns are consistent with findings reported in the literature? Consider Bresnan and Hay (2008), who find – based on many of the same dative tokens that feed into the NZE/AmE sub-datasets under study here – that non-animate recipients are more likely to be used in the prepositional dative construction in spoken NZE than in spoken AmE (Bresnan & Hay 2008: Figure 1). Our analysis fails to replicate this effect, but we note that our data were coded differently,[9] and that we do not, in contrast to Bresnan and Hay, consider information status in the model. Moving on in the literature, we note that Wolk et al. (2013) find that theme length effects have a stronger effect in AmE than in BrE. We do not see this effect in our data, but we would like to caution that Wolk et al. studied written and historical data while we study contemporary spoken data. Tagliamonte (2014) reports two differences between CanE and BrE: first, a greater use of the ditransitive dative in BrE (a finding which we replicate), and second, a greater degree of spread of the ditransitive dative into animate recipients in BrE compared to CanE, where the prepositional dative was still robust up to the more recent generations. Our analysis was unable to confirm significant differences with regard to recipient animacy; instead we find two contrasts not reported by Tagliamonte (2014). It should be borne in mind however that the modeling techniques we used differ substantially from those employed in Tagliamonte (2014). How replicable are these findings? The BrE/CanE, BrE/NZE, and CanE/NZE contrasts detailed in Table 11 are not replicated in Röthlisberger et al. (to appear), a study on the probabilistic grammar of the dative alternation that covers, among others, the same varieties that are represented in Table 11 (however, the regression models are designed differently and the data sources subject to study – the International Corpus of English and the Corpus of Global Web-Based English – are more acrolectal and less vernacular than the data we study here).

6 Discussion and directions for future research

In this paper we introduced two datasets designed to facilitate the comparative investigation of spoken syntax. Based on this new resource, we investigated the extent to which patterns in the English genitive and dative alternations differ across four major regional varieties of English. Drawing inspiration from work in Probabilistic Grammar (Bresnan & Ford 2010), variationist (socio)linguistics (Labov 1982) and – more specifically – comparative sociolinguistics (Tagliamonte 2001), we took a closer look at the extent to which constraints on genitive and dative choice are fluid or stable in a cross-variety perspective.

Our key findings may be summarized as follows. We utilized conditional random forest analysis (CRF) to probe the overall importance of constraints, and found that the constraint rankings looked more similar across varieties in the genitive alternation (where possessor animacy is consistently the most important constraint by far) than for the dative alternation, where there is more cross-varietal fluidity. Subsequently, we used binary logistic regression analysis with mixed effects to investigate the effect size and significance of the constraints on dative and genitive variation. We specifically conducted careful pairwise regression comparisons, in which we were interested in potential interactions between variety and language-internal constraints. We found five such pairwise probabilistic contrasts in the genitive alternation, and eight in the dative alternation. Analogous to the CRF analysis, therefore, we observe that with regard to the sheer number of probabilistic contrasts the genitive alternation is more uniform in a cross-variety perspective than the dative alternation. In addition, we noted that some constraints on the dative alternation (semantics of the dative verb *give*) have different effect directions in different varieties, a sort of reversal that we do not see in the genitive alternation.

All things considered, we are struck by how similar and homogeneous the probabilistic grammars are across the spoken varieties under study in this paper, despite more or less

[9] In particular, we note that a recently-discovered and previously-unreported error saw Bresnan and Hay (2008) code organizations as "animate" in the US data but "inanimate" in the NZ data. Since constructions with an organization as possessor tend to be realized as *s*-genitives less often than constructions with a (non-organization) animate possessor, this difference in coding may explain why AmE speakers were found to be less likely than NZE speakers to use the *s*-genitive with a possessor coded as "animate".

subtle differences between the component sub-datasets on which our analysis is based. To highlight this homogeneity further and to provide some quantitative context for the discussion of the contrasting effect(s) later on, we take the liberty now to report two supplementary and deliberately very simple fixed-effects binary logistic regression models, one per alternation (see Tables 12 and 13).[10] Crucially, the models do not include any variety-related information, either as main effect or in interaction terms, and are thus

[10] The predictors in these models are the constraints discussed in Section 3. The models were calculated using R's glm() function. The models were not optimized in any way.

Table 12: Constraints on genitive variation in fixed-effects logistic regression analysis across all varieties. Predicted odds are for the *s*-genitive.

	coefficient (*b*)	
(Intercept)	–2.11	***
possessor animacy (default: inanimate)		
animate	5.47	***
collective	2.18	***
locative	1.34	***
temporal	1.71	***
possessor definiteness (default: definite)		
definite proper noun	0.04	
indefinite	–0.65	***
prototypical semantics (default: non-prototypical)	1.51	***
possessor length	–0.85	***
possessum length	0.17	*
final sibilancy (default: no final sibilancy)	–1.18	***

C = 0. 96. Significance codes: "***": $p < 0.001$; "**": $p < 0.01$; "*": $p < 0.05$. Positive coefficients favor the *s*-genitive; negative coefficients disfavor.

Table 13: Constraints on dative variation in fixed-effects logistic regression analysis across all varieties. Predicted odds are for the prepositional dative.

	coefficient (*b*)	
(Intercept)	–4.03	***
semantics of the dative verb (default: transfer)		
abstract	–0.59	**
communication	–0.62	*
length difference recipient/theme	1.58	***
recipient type: non-pronominal (default: pronominal)	2.11	***
theme type: pronominal (default: non-pronominal)	2.44	***
recipient definiteness: indefinite (default: definite)	1.30	***
theme definiteness: definite (default: indefinite)	2.12	***
recipient animacy: inanimate (default: animate)	1.75	***
theme animacy: animate (default: inanimate)	0.18	

C = 0.96. Significance codes: "***": $p < 0.001$; "**": $p < 0.01$; "*": $p < 0.05$. Positive coefficients favor the prepositional dative; negative coefficients disfavor.

entirely agnostic about variety differences. Nonetheless, most of the predictors are significant and have the theoretically expected effect directions (and the models have excellent fits). The point is that the parallel main effects of the constraints across the varieties can be interpreted as evidence for a solid probabilistic core grammar. In other words, while comparative analysis and the quest for contrasts took center stage in this study, probabilistic homogeneity seems to be the key trait of the datasets under study.

Against the backdrop of this overall homogeneity, we did identify several significant probabilistic contrasts (Tables 4 and 11), and the existence of such contrasts is expected if grammar indeed constitutes "the cognitive organization of one's experience with language" (Bybee 2006: 711), because such experiences vary across communities. For example, one of the constraints that is malleable across varieties is final sibilancy in the genitive alternation: the effect of final sibilancy is stronger in NZE than in BrE and CanE (see Table 4). It is interesting to note in this context that the avoidance of a clash in final sibilancy is a predictor of the genitive but not the dative alternation: the same speakers who use the *of*-genitive to avoid a possessor's final sibilant in the *s*-genitive, do not also avoid the ditransitive dative when the recipient has a final sibilant adjacent to an initial sibilant in the theme (Bresnan 2011). Avoidance of the final sibilant thus appears to be a (probabilistic) grammatical property of the genitive construction, which cannot be dismissed as a product of general articulatory constraints on adjacent sibilants. Hence our finding in the present study that some varieties are reliably differentiated by the likelihood of final-sibilance avoidance in the genitive is very strong evidence for probabilistic grammars of these varieties.

But in all, the number of probabilistic contrasts we observed is relatively small, taking into account the numerous comparisons between sub-datasets differing in geographical origin, speaker attributes, and collection methods. The degree of cross-varietal homogeneity as regards the strengths and directions of the various constraints is rather remarkable: in the big picture, it is the parallels across varieties that really stand out, and so we suggest the findings here are one further step toward identifying a model of syntactic variation that forms part of the "common core" (Quirk et al. 1985: 16) of the grammar of spoken English.

The homogeneity (again, despite some heterogeneity in the materials) we see in the models shows a common statistical pattern, termed "quantitative harmonic alignment" (Aissen 1999; Bresnan et al. 2007):

> One of the main findings of previous corpus work on the dative alternation is the existence of a statistical pattern in which, all else being equal, animate, definite, pronominal, discourse-accessible, and shorter arguments tend to precede inanimate, indefinite, nonpronominal, less discourse-accessible, or longer arguments in both of the dative constructions. (Bresnan & Ford 2010: 181)

Likewise in the genitive alternation, animate, definite, pronominal, discourse-accessible, and shorter genitive constituents should precede inanimate, indefinite, nonpronominal, less discourse-accessible, or longer genitive constituents (Rosenbach 2014: 238–239). What we are seeing in our analysis is that the basic common phenomena of quantitative harmonic alignment are very robust indeed across the varieties, even as we have looked hard for probabilistic differences between varieties.

We were unable to find some contrasts in our speech-only materials that have been reported for written varieties (e.g. that long possessums favor s-genitive usage in AmE but not in BrE; see Hinrichs & Szmrecsanyi 2007). The reason for this, then, may well be that spoken vernaculars of the type covered in our datasets are more homogeneous than written varieties. This may seem surprising to those who would suppose that spoken language should be more heterogeneous in nature than written language, for which stylistic norms (which may be regional, however) seem to be in place. But it is probably not entirely unreasonable to assume that the production of spoken syntax – we emphasize that this study is not about lexis or accent differences – is subject to processing and production constraints and biases (e.g. Hawkins 1994; MacDonald 2013) in a way that the production of written syntax is not; hence the homogeneity. That being said, we believe that the cross-variety differences that we do find are consistent with the predictions of a framework assuming that subtle usage differences are learnt statistically. We know, for example, that "processing" concepts such as "working memory" in comprehension tasks are affected by training (i.e. experience) (see, e.g., Wells et al. 2009). Broadly viewed, then, statistical learning is a usage-based and/or experience-based framework, and the probabilistic contrasts that our modeling has revealed, although not exceedingly numerous, are consistent with predictions generated by this framework.

Contrary to our expectations, we also did not find many of the cross-variety contrasts reported in recent studies of spoken language (e.g. the animacy-related AmE/NZE contrast reported in Bresnan & Hay 2008). At the same time, we mentioned how recent follow-up research (Heller et al. 2017; Röthlisberger et al. To appear) fails to replicate contrasts that we have uncovered in this study. These replication failures are potentially troubling – it seems that comparative variation analysis of the type we have performed here is more sensitive to the particulars of the data source and coding decisions and to the design of the analysis than one would suspect or hope. In this regard, however, our empirical work makes a valuable contribution; our careful scrutiny of inconsistencies in coding decisions within and across varieties and constructions, by multiple authors, has yielded datasets that are both more principled and more comparable than those used in previous studies. Consequently, while the differences between previous and present results call into question the extent to which any set of results can be generalized, we remain confident that the picture gained from the present datasets is more accurate than the picture gained from previous datasets.

Nevertheless, the degree to which probabilistic contrasts are robust across data sources and research designs needs more attention. For one thing, we conducted multiple pairwise comparisons but – in a rather exploratory spirit – did not adjust the significance threshold accordingly. Thus (except for final sibilancy in the genitive alternation) many of our interaction effects do not reach Bonferroni-corrected thresholds for significance ($p = .05/6 = .008$), and so it is maybe not too surprising that many of the contrasts that we describe here fail to replicate in other studies. In general, then, it may be the case that the datasets we investigate are too small, and the predictor sets we are using too large, for regression analysis to reliably pick up and replicate probabilistic contrasts. Other analysis techniques, such as conditional inference trees (see also below), may be better suited to explore "small *n* and large *p*" datasets such as ours. Another issue that should be kept in mind in this connection is that we have taken the liberty in this study to not consider social factors (for which the datasets are partly annotated), but it is clear that in doing so we may have introduced confounds. Therefore, including sociolinguistic predictors in future analysis is high on the agenda.

Other directions for future research include the following. It would be highly desirable to provide a systematic comparison of different ways to model the data; these need not necessarily be regression-based, as in this paper, but may also include less widely used techniques such as memory-based learning (Daelemans & Bosch 2005) or naïve discriminative learning (Baayen 2011). As for designing models of the genitive alternation specifically, with possessor animacy being such a crucial (and in some cases near-categorical) constraint it may be worth considering taking possessor animacy out of the regression models at a later step in the analysis in order to zoom in on the attributes of the variation grammars that are actually divergent and highly variable (Tagliamonte 2014 is a recent study that uses this technique; the idea goes back to Labov 1969: 729, who argues that inclusion of (near-)categorical contexts will obscure the real patterns of variation). Alternatively, it may be worthwhile to turn to conditional inference trees (e.g. Hothorn, Hornik & Zeileis 2006), an analysis technique which is crucially based on this sort of data sub-setting. Lastly, future research should also study the possibility that the (subtle) differences across varieties are due to differential stages in the evolution of the syntactic variants themselves: the datasets on which the present study's analysis is based are an excellent resource to investigate how speakers of particular vernacular varieties are slower or quicker to adopt progressive syntactic variants in particular contexts (see Tagliamonte, Durham & Smith 2014).

Appendices

The datasets on which the empirical analysis is based can be freely downloaded, along with documentation, from https://purl.stanford.edu/qj187zs3852.

Abbreviations

BrE = British English, CRF = Conditional Random Forest analysis, ONZE = Canterbury Corpus of the Origins of New Zealand English, CSAE = Corpus of Spoken American English, CanE = Canadian English, FRED = Freiburg English Dialect Corpus, NZE = New Zealand English, ODA = Ontario Dialects Archive, AmE = US American English.

Supplementary File

The supplementary file for this article can be found as follows:

- Supplement to 'Spoken syntax in a comparative perspective'. DOI: https://doi.org/10.5334/gjgl.310.s1

Acknowledgements

Valuable feedback from the audience of the 2016 workshop on "Probabilistic variation across dialects and varieties", held at KU Leuven, is gratefully acknowledged. Thanks go to Melanie Röthlisberger and Benedikt Heller for help with fitting the regression models. The usual disclaimers apply. This material is based upon work supported by the National Science Foundation under Grant No. BCS-1025602, and the Research Foundation – Flanders (FWO) under Grant No. G.0C59.13N.

Competing Interests

The authors have no competing interests to declare.

References

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language & Linguistic Theory* 17(4). 673–711. DOI: https://doi.org/10.1023/A:1006335629372

Allen, Cynthia L. 2006. Case syncretism and word order change. In Ans van Kemenade & Bettelou Los (eds.), *The handbook of the history of English*, 201–223. Chichester: Wiley-Blackwell. DOI: https://doi.org/10.1002/9780470757048.ch9

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511801686

Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295–328. DOI: https://doi.org/10.1590/S1984-63982011000200003
Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker. 2015. Fitting linear mixed effect models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01
Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142.
Breejen, Bastian den. 1937. The Genitive and its "of"-equivalent in the latter half of the 16th century. Amsterdam: Universiteit Amsterdam.
Breiman, Leo. 2001. Random forests. *Machine Learning* 45. 5–32. DOI: https://doi.org/10.1023/A:1010933404324
Bresnan, Joan. 2011. Acquiring syntactic variation in English: a cross-constructional study. Plenary talk presented at the 6th International Conference on Language Variation in Europe (ICLAVE 6), Freiburg.
Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Boume, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245–259. DOI: https://doi.org/10.1016/j.lingua.2007.02.007
Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213. DOI: https://doi.org/10.1353/lan.0.0189
Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733. DOI: https://doi.org/10.1353/lan.2006.0186
Clifton, Charles & Lyn Frazier. 2004. Should given information come before new? Yes and no. *Memory & Cognition* 32(6). 886–895. DOI: https://doi.org/10.3758/BF03196867
Daelemans, Walter & Antal van den Bosch. 2005. *Memory-based language processing* (Studies in Natural Language Processing). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511486579
De Cuypere, Ludovic. 2010. A discourse-based account of the Old English double object alternation. *Sprachwissenschaft* 35. 337–368.
Eitelmann, Matthias. 2016. Support for end-weight as a determinant of linguistic variation and change. *English Language and Linguistics* 20(03). 395–420. DOI: https://doi.org/10.1017/S1360674316000356
Fischer, Olga. 1992. Syntax. In Norman Blake (ed.), *The Cambridge history of the English language* 2(1066–1476). 207–408. Cambridge: Cambridge University Press.
Fischer, Olga & Wim van der Wurff. 2006. Syntax. In Richard Hogg & David Denison (eds.), *A history of the English language*, 109–198. Cambridge: Cambridge University Press.
Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models* (Analytical Methods for Social Research). Cambridge: Cambridge University Press.
Godfrey, John J, Edward C. Holliman & Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *IEEE International Conference on Speech, and Signal Processing, ICASSP-92* 1. 517–520. DOI: https://doi.org/10.1109/ICASSP.1992.225858
Gordon, Elizabeth, Margaret Maclagan & Jennifer Hay. 2007. The ONZE corpus. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and digitizing language*

corpora 2. 82–104. Diachronic Databases, Basingstroke: Palgrave Macmillan. DOI: https://doi.org/10.1057/9780230223202_4

Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(03). 471–496. DOI: https://doi.org/10.1017/S1360674314000136

Harrell, Frank E. 2001. *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer New York. DOI: https://doi.org/10.1007/978-1-4757-3462-1

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

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. DOI: https://doi.org/10.1177/0075424216685405

Hernández, Nuria. 2006. User's guide to FRED. www.freidok.uni-freiburg.de/volltexte/2489. Freiburg: English Dialects Research Group.

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. 437–474. DOI: https://doi.org/10.1017/S1360674307002341

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

Hundt, Marianne & Benedikt Szmrecsanyi. 2012. Animacy in early New Zealand English. *English World-Wide* 33. 241–263. DOI: https://doi.org/10.1075/eww.33.3.01hun

Jahr Sohrheim, Mette. 1980. *The s-genitive in present-day English*. Oslo: University of Oslo dissertation.

Jankowski, Bridget L. & Sali A. Tagliamonte. 2014. On the genitive's trail: Data and method from a sociolinguistic perspective. *English Language and Linguistics* 18(02). 305–329. DOI: https://doi.org/10.1017/S1360674314000045

Jucker, Andreas. 1993. The genitive versus the *of*-construction in newspaper language. In Andreas Jucker (ed.), *The noun phrase in English. Its structure and variability*, 121–136. Heidelberg: Carl Winter.

Kachru, Braj B. (ed.). 1992. *The other tongue: English across cultures* (English in the Global Context). 2nd edition; Urbana, IL: University of Illinois Press.

Kemenade, Ans van. 1987. *Syntactic case and morphological case in the history of English*. Dordrecht & Providence, RI: Foris. DOI: https://doi.org/10.1515/9783110882308

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

Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), *Perspectives on historical linguistics*, 17–92. Amsterdam: Benjamins. DOI: https://doi.org/10.1075/cilt.24.06lab

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

Mair, Christian. 2013. The world system of Englishes: accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34(3). 253–278. DOI: https://doi.org/10.1075/eww.34.3.01mai

Mitchell, Bruce. 1985. *Old English syntax*. Oxford & New York, NY: Clarendon Press. DOI: https://doi.org/10.1093/acprof:oso/9780198119357.001.0001

Mustanoja, Tauno F. 1960. *A Middle English syntax* I. Helsinki: Société Néophilologique.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London & New York, NY: Longman.
R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.
Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin & New York, NY: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110899818
Rosenbach, Anette. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81(3). 613–644. DOI: https://doi.org/10.1353/lan.2005.0149
Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18(02). 215–262. DOI: https://doi.org/10.1017/S1360674314000021
Rosenbach, Anette. 2017. Constraints in contact: Animacy in English and Afrikaans genitive variation – a cross-linguistic perspective. *Glossa: a journal of general linguistics* 2(1). 72. DOI: https://doi.org/10.5334/gjgl.292
Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. To appear. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics*.
Schneider, Edgar. 2007. *Postcolonial English: varieties around the world*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511618901
Shih, Stephanie, Jason Grafmiller, Richard Futrell & Joan Bresnan. 2015. Rhythm's role in genitive construction choice in spoken English. In Ralf Vogel & Ruben Vijver (eds.), *Rhythm in cognition and grammar*. Berlin, München & Boston, MA: De Gruyter. DOI: https://doi.org/10.1515/9783110378092.207
Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(1). 307. DOI: https://doi.org/10.1186/1471-2105-9-307
Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348. DOI: https://doi.org/10.1037/a0016973
Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English*. Berlin & New York, NY: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110197808
Szmrecsanyi, Benedikt. 2016. About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory* 12(1). DOI: https://doi.org/10.1515/cllt-2015-0068
Szmrecsanyi, Benedikt, Anette Rosenbach, Joan Bresnan & Christoph Wolk. 2014. Culturally conditioned language change? A multivariate analysis of genitive constructions in ARCHER. In Marianne Hundt (ed.), *Late Modern English syntax*, 133–152 (Studies in English Language). Cambridge: Cambridge University Press.
Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: Benjamins. DOI: https://doi.org/10.1075/silv.2.22szm
Tagliamonte, Sali. 2001. Comparative sociolinguistics. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *Handbook of language variation and change*, 729–763. Malden, MA & Oxford: Blackwell.

Tagliamonte, Sali A. 2014. A comparative sociolinguistic analysis of the dative alternation. In Rena Torres-Cacoullos, Nathalie Dion & André Lapierre (eds.), *Linguistic variation: Confronting fact and theory*, 297–318. London & New York, NY: Routledge.

Tagliamonte, Sali A. & 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. 135–178. DOI: https://doi.org/10.1017/S0954394512000129

Tagliamonte, Sali A., Mercedes Durham & Jennifer Smith. 2014. Grammaticalization at an early stage: Future be going to in conservative British dialects. *English Language and Linguistics* 18(01). 75–108. DOI: https://doi.org/10.1017/S1360674313000282

Thomas, Russel. 1931. *Syntactical processes involved in the development of the adnominal periphrastic genitive in the English language*. Ann Arbor, MI: University of Michigan dissertation.

Traugott, Elizabeth Closs. 1992. Syntax. In Richard Hogg (ed.), *The Cambridge history of the English language* 1. 168–289 The beginnings to 1066. Cambridge: Cambridge University Press.

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

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

Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. New York, NY: Springer. DOI: https://doi.org/10.1007/978-0-387-87458-6

How to cite this article: Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte and Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2(1): 86. 1–27, DOI: https://doi.org/10.5334/gjgl.310

Submitted: 07 December 2016 **Accepted:** 27 Jun 2017 **Published:** 29 September 2017

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press. **OPEN ACCESS**