Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice

Sali A. Tagliamonte and R. Harald Baayen
University of Toronto & University of Tübingen and University of Alberta

February 2012

Short title: *Was/were* as a case study for statistical practice

Contact information for lead author:

Department of Linguistics
University of Toronto
100 St George Street,
Sid Smith HallRoom 4077
Toronto, Ontario M5S 3H1
Canada
sali.tagliamonte@utoronto.ca
Tel: 416 946-8024
Fax: 416 971-2688

Abstract

What is the explanation for vigorous variation between *was* and *were* in plural existential constructions and what is the optimal tool for analyzing it? Previous studies of this phenomenon have used the the variable rule program, a generalized linear model; however, recent developments in statistics have introduced new tools, including mixed-effects models, random forests and conditional inference trees which may open additional possibilities for data exploration, analysis and interpretation. In a step-by-step demonstration, we show how this well known variable benefits from these complementary techniques. Mixed-effects models provide a principled way of assessing the importance of random-effect factors such as the individuals in the sample. Random forests provide information about the importance of predictors, whether factorial or continuous, and do so also for unbalanced designs with high multicollinearity, cases for which the family of linear models is less appropriate. Conditional inference trees straightforwardly visualize how multiple predictors operate in tandem. Taken together the results confirm that polarity, distance from verb to plural element and the nature of the DP are significant predictors. Ongoing linguistic change and social reallocation via morphologization are operational. Furthermore, the results make predictions that can be tested in future research. We conclude that variationist research can be substantially enriched by an expanded tool kit.

1 Introduction

The choice of optimum statistical tool for analyzing linguistic variation has a long history of controversy in quantitative sociolinguistics, beginning from the role of statistics in the study of variation (e.g., Bickerton, 1971, 1973; Kay, 1978; Kay and McDaniel, 1979; Downes, 1984) and continuing on to controversies over the the application of statistical methods to morpho-syntactic variables (e.g., Rickford, 1975; Lavandera, 1978) and discourse-pragmatic variables in the 2000's (e.g., Cheshire, 2005). Currently, the debate centers not on whether statistical methods are appropriate, but on the choice of which one is the best. The variable rule program, in its various guises as *Varbrul* (Cedergren and Sankoff, 1974), *Goldvarb* 2.0 (Rand and David Sankoff, 1990), *Goldvarb X* (Sankoff, 2005), or *Goldvarb Lion* (Sankoff, Tagliamonte, and Smith, 2012), is a particular implementation of the generalized linear model for data that have two discrete variants (i.e. binary count data). It is capable of modelling the joint effect of many independent (orthogonal) factors. General statistical packages such as SAS, SPSS and R offer comparable models.

However, developments in statistics over the past 30 years have introduced new statistical techniques, including generalized mixed-effects models, capable of modeling subtle differences among internal and external factors (e.g., Bates, 2005; Baayen, 2008; Baayen, Davidson, and Bates, 2008; Jaeger, 2008; Johnson, 2009). Such models have come into language variation and change studies through statistical packages as *Rvarb* (Paolillo, 2002), *Rbrul* (Johnson, 2009), and *R* (R Development Core Team, 2009). However, many researchers in language variation and change do not understand the differences among these statistical packages and the tools they offer, nor do they have the background to make informed decisions about how to use different models most effectively. Indeed, the 'tool', the generalized linear model vs. the generalized linear mixed model, is often confused with the 'toolkit', namely *Goldvarb* vs. SPSS, SAS, or R.

In this paper, our quest to further understand *was/were* variation will lead us to explore some new tools on the market, focussing on the concepts and ideas that make them useful to language variation and change analysts more generally. One such tool, the generalized linear mixed model, is implemented in many general software packages, both commercial (SPSS, SAS) and open-source (R), as well as in more specialist software (e.g. MLwiN, 2007; Gilmour et al., 2002). For a cross-platform guide to mixed modeling, see West, Welch, and Galecki (2007). We also discuss a more recent tool, known as random forest and bagging ensemble algorithms, a relatively recent and novel type of non-parametric data analysis. Non-parametic analyses make no assumptions about the distribution of the population from which a sample was drawn (e.g. Baayen; 2008:77). The implementation that we have used (which uses conditional inference trees, see Strobl, Boulesteix, Zeileis, and Hothorn, 2007; Strobl, Boulesteix, Kneib, Augustin, and Zeileis, 2008; Hothorn, Hornik, and Zeileis, 2006b) is, to the extent of our knowledge, only available in R. The appendix provides the reader with the R code for replicating the analyses reported here. To facilitate our ancillary goal of introducing new methods of analysis, the data are available on the first author's website (`http://individual.utoronto.ca/tagliamonte/Downloads/york.csv`).

Our aim is to demonstrate that these new tools enrich the variationist toolkit by offering new ways for understanding language variation. The study of *was/were* variation using different types of statistical analyses lead us to a number of new conclusions about the variable.[1]

[1]This paper grew out of a discussion at a workshop held at NWAV 38 in Ottawa, Canada in October 2009 entitled Using Statistical Tools to Explain Linguistic Variation (Tagliamonte, 2009). The workshop brought together leading proponents of a range of different statistical tools and methods in order to exchange views. This final version of the paper benefited from the critical eye of three astute LVC reviewers as well as detailed comments from Alexandra D'Arcy. We thank everyone for their input.

2 Was/were variation

Variation between *was* and *were* in past tense plural existential constructions, as in (1) can be found in virtually any data set, in any variety of English, in any location in the world. The data upon which we base our analyses come from the city of York in northeast England from the late 1990's where the following examples are typical of this linguistic variable in conversation.

(1) a. There *wasn't* the major sort of bombings and stuff like that but there *was* orange men. You know there /em was odd things going on. (YRK/074)

b. There *was* one or two killed in that area, and um, we think- ... we think there *were* firemen killed. (YRK/022)

This linguistic variable has been studied in varieties of English spanning the globe, including the United States, Canada, United Kingdom, Australia, New Zealand and various places in between (e.g., Britain, 2002; Britain and Sudbury, 1999; Cheshire, 1982; Christian, Wolfram, and Dube, 1988; Trudgill, 1990; Eisikovits, 1991; Hay and Schreier, 2004; Hazen, 1996; Meechan and Foley, 1994; Milroy and Milroy, 1993; Montgomery, 1989; Schreier, 2002; Tagliamonte and Smith, 1998, 2000; Trudgill, 1990; Walker, 2007). Indeed, fundamental developments within language variation and change studies from the 1960's through to the present have come from analyses of this linguistic feature (e.g., Fasold, 1969, 1972; Labov, 1969; Labov, Cohen, Robins, and Lewis, 1968; Wolfram, 1969; Wolfram and Christian, 1976). Moreover, this phenomenon plays a key role in ongoing explorations of the relationship between linguistic variation and linguistic theory, i.e. socio-syntax (e.g., Adger, 2006; Adger and Smith, 2005, 2007; Cornips and Corrigan, 2005), of processing effects in psycholinguistics (e.g., Bock and Miller, 1991; Bock and Kroch, 1988) and of refinements to theoretical models of language (e.g. Biberauer and Richards, 2008; Börgars and Chapman, 1998; Henry, 1995, 1998; Meechan and Foley, 1994), making it a key variable in the history of the discipline. Yet, despite the remarkably broad and extensive information base on *was/were* variation, there are still conflicting explanations for its role and function in the grammar of English. One of the reasons for this state of affairs is the complexity of the data, which gives rise to many problems for statistical analysis.

3 The data

In this paper, we focus on *was/were* variation in its most ubiquitous context[2] — past tense plural existential constructions. The particular question we begin with is what explains *was/were* variation? Although typically viewed as a universal of vernacular English, an enriched statistical toolkit will enable us to probe the question why. Beginning with the results of a variable rule analysis Sankoff (1988), we then show how generalized mixed-effects modeling and modeling with the help of random forests can lead to a more nuanced understanding of the phenomenon.

In a 1998 study of York English, all past tense contexts of the verb 'to be' were examined totaling nearly 7000 tokens from 40 different individuals. Use of *was* was generally receding across all areas of the grammatical paradigm, yet in plural existential constructions it was very frequent (Tagliamonte, 1998, , 181, Table 12). Separate analysis of the 310 plural existential tokens in the sample suggested that in this context non-standard *was* was increasing in apparent time. Further, its use appeared to be the result of internal syntactic relations; however, this result was never fully explored (Tagliamonte, 1998, 186). Given the building body of evidence from a decade more

[2]Present tense existentials are also frequent and widespread. However the status of *There's* as a grammaticalized or fused collocate in the language may obscure grammatical patterning (Walker, 2007, p.160-162).

of extensive study of *was/were* variation in plural existential constructions, now encompassing innumerable speech communities, dialects and localities as well as many different perspectives from different areas of the discipline, there is a considerably deeper knowledge base and understanding with which to re-examine the York materials, exploit the full data set of 91 individuals and dig deeper into the data on existentials.

The York English corpus provides a relatively large spoken language corpus that is socially stratified and informal, and which represents the spoken English of a particular time and place in the history of the language. The variety of English spoken in the city is a northern variety of British English. Although it retains a number of local features (Tagliamonte and Roeder, 2009); it has been previously shown to be participating in many current changes in progress in the United Kingdom and in English generally (Tagliamonte, 2001, 2002a,b, 2003; Tagliamonte and Smith, 2006). Thus, it offers a view on the typical patterns of English usage at the turn of the 21st century. In the present study, the corpus was exhaustively searched for all plural past tense existential constructions. Each context retained in the analysis was coded for the most prominent set of factors extrapolated from the historical and contemporary literature on *was/were* variation in existentials. In the analyses that follows, we include the major effects previously reported, both social (age, sex and education) and linguistic (polarity, type of determination and proximity of the verb (*was* or *were*) to its referent). In addition, we especially scrutinize the contrast between categorical and variable individuals as well as the nature of the link between verb and referent.

3.1 External factors

Was/were variation has typically demonstrated sociolinguistic patterns undoubtedly due to the fact that the two variants are clearly split between standard (i.e. *were*) and non-standard (i.e. *was*). Not surprisingly, most studies report socio-economic effects: *Was* is more frequent among working class individuals and formality increases the use of the standard form (Christian et al., 1988; de Wolf, 1990; Eisikovits, 1991; Feagin, 1979; Hay and Schreier, 2004; Schreier, 2002; Tagliamonte and Smith, 2000). However, there is irregularity in the findings across studies for sex differentiation and this varies depending on the time depth of the data (see Hay and Schreier, 2004, Figure 1, p. 216). Given the non-standard status of plural existential *was*, the expectation is that males will favour this form. However, in many studies females are the more frequent users. Moreover, the intersection of age and sex is widely held to be a key component of the suite of explanatory factors. Thus, it could be the case that interactions between social predictors have not been sufficiently accounted for in earlier studies. This would account for the inconsistent results. Nevertheless, a number of studies have noted increasing frequency of *was* for younger people and that women lead in this linguistic development, e.g. Appalachian English (Montgomery, 1989), Tristan da Cunha English (Schreier, 2002), New Zealand English (Hay and Schreier, 2004), Australia English (Eisikovits, 1991). Thus, while *was/were* variation might appear to be a classic sociolinguistic variable, there are indications of widespread ongoing change in progress. This leads to a (partial) reason why this variation remains robust in contemporary varieties. Despite being socially stigmatized, apparently there is a more universal change in progress. While this explanation is attractive, it is critical to point out that all previous studies have treated age as a factorial predictor which essentially breaks the individuals into age groups. Thus, it could be the case that unaccounted individual differences underlie the interpretation of ongoing change. In sum, there is still no full explanation for why *was/were* variation is so productive or how it may be evolving in contemporary English if at all. This calls for a re-examination of the variable with an enhanced data set and an enriched analytic toolkit.

In keeping with our goal to promote a bridge from sociolinguistic practice to 'general statistical practice', we will use standard statistical terminology. We use the term *predictor* as a term covering both numerical predictors (or covariates), and factorial predictors and we refer to the values of a factorial predictor as its *levels*. In sociolinguistic practice, a factorial predictor would usually be referred to as 'factor group' or 'factor'. In what follows, we represent factor names in typewriter font, and put the levels of a factor in italics.

In our analyses, the external factors are represented by the following predictors: `Sex` (a factor with levels *male* and *female*), `Education` (a factor with levels `low` and `high`), and `Age` or `AgeGroup`. `AgeGroup` is a factor predictor with levels *20–30, 31–50, 51–70*, and *70+*, whereas `Age` is a numeric predictor with the age in years of the individual, e.g. 21, 26, 35, 39, 53, 62 etc. We will compare results for age when treated as a numeric rather than factor predictor.

3.2 Internal factors

Perhaps the most widely attested linguistic constraint on *was/were* variation is the effect of polarity, the contrast between affirmative and negative (e.g. Anderwald, 2002; Britain and Sudbury, 1999; Britain, 2002; Schilling-Estes and Wolfram, 1994; Hazen, 1996). This effect is present across most varieties, but differs in nature from one variety to the next. The pattern can be explained as a realignment of the past tense morphological forms *was* and *were* towards a contrast between negative and affirmative contexts. The most widespread version of this effect is the case where *weren't* occurs more often in negatives and *was* occurs more often in affirmatives. This was the pattern found in the earlier analysis of York (Tagliamonte, 1998, 180), as in (2).

(2) There *weren't* always bulls. Sometimes there *was* a few pigs, a few sheep . . . (YRK/002)

The same pattern is attested in some North American dialects in the US, e.g. North Carolina, (Schilling-Estes and Wolfram, 1994) and across England, including the southwest (Reading) (Cheshire et al., 1995), the Fens in the southeast (Britain and Sudbury, 2002) and elsewhere in Britain (Anderwald, 2002).

The second pattern is when the contrast goes in the opposite direction: non-standard *was* occurs more often with negatives, i.e. *wasn't*, and the standard form *weren't* occurs with affirmatives. This pattern is found in other northern Englishes, e.g. Northern England (Maryport), southwest Scotland (Cumnock) and Northern Ireland (Portavogie and Cullybackey) (Tagliamonte, 2009) (Tagliamonte and Smith, 2000, 160–161), as in (3).

(3) a. There *wasn't* any fancy puddings nor no fancy cake nor biscuits. (CMK/-)
b. There *were* a whole lot of them. (CMK/-)

This pattern is reported for in North America for varieties such as African Nova Scotian English (Tagliamonte and Smith, 2000).

A third pattern is also attested. This is where the *was* variant occurs regardless of polarity, i.e. no polarity effect. This is the pattern reported for New Zealand English, apparently across the 18th and 19th centuries (Hay and Schreier, 2004, p.228) and (Chambers, 2004, p. 131), as exemplified in (4).

(4) No, there *wasn't* too many cars. There *was* some, but there *wasn't* a great many. (WHL/S)

Thus, remorphologization, a common process whereby syntactic phenomena develop morphological contrasts (Joseph and Janda, 1986, 2003) often appears to be an underlying explanation for *was/were* variation in existentials (see Schilling-Estes and Wolfram, 1994).

Polarity is coded as a factor named `Polarity` with as levels *Affirmative* and *Negative*.

Another prominent constraint on this variable relates to the proximity of the verb to its referent or to a plural element (Britain, 2002; Hay and Schreier, 2004; Tagliamonte, 1998). These effects have been subsumed under various labels, including 'number shift', 'attraction', 'confusion of proximity' among others. However, they may be explained by different underlying phenomena. One hypothesis is that the agreement relationship between verb and referent becomes compromised when they are distant from each other. This distance is hypothesized to hamper agreement, either as a barrier to direct Case assignment (e.g., Henry, 1995) or for more general processing reasons (Bock and Kroch, 1988). Thus, a crucial consideration is the nature of this underlying relationship. Another hypothesis predicts that the form of the verb will be influenced by a close plural element. Depending on the underlying hypothesis, this predictor must be categorized quite differently. The examples in (5a–l) will serve to illustrate this.

(5) a. There *were* badgers in there. (YRK/087)
  b. There *was* [black] clouds. (YRK/078)
  c. There *were* [two] pubs. (YRK/012)
  d. There *were* [the] eggs. (YRK/031)
  e. There *was* [no] treats for them. (YRK/042)
  f. There *was* [some funny] people. (YRK/048)
  g. There *was* [all little] houses in there. (YRK/011)
  h. There *was* [lots of] cinemas. (YRK/16)
  i. There *was* [always] two films on. (YRK/003)
  j. There *was* [four of these] houses ... (YRK/048)
  k. There *was* [about twelve different] groups ... (YRK/077)
  l. There *was* [still quite strong] winds in this part. (YRK/078)
  m. There *was* [like purple and green and yellow] bruises. (YRK/049)

For proximity, we have coded the following predictors. First, `Proximate1` assesses the number of words intervening between verb and plural element. For example, in (5c) the verb and the first plural element, *two*, are adjacent whereas in (5k) there is one word intervening, the adverb *about* intervenes between *was* and *twelve*. As the counts of the numbers of intervening words vary substantially (from 1 for 6 intervening words to 198 for 1 intervening word), we also considered a binary factor `Proximate1.adj` with as levels *Adjacent* (171 observations) and *Non-Adjacent* (318 observations), contrasting all cases of adjacent verb to plural element sequences vs. non-adjacent ones.

`Proximate2` assesses the number of words intervening, but in this configuration, it is the relationship between verb and referent that is relevant. For example, (5h) has two words while (5m) has six words intervening. `Proximate1` and `Proximate2` distinguish the position of the referent NP vs. a pluralizing element by number of words. As the counts for the different distances vary substantially (from 1 for distances 7 and 8, to 130 for distance 1), we also introduced a factor, labelled `Adjacency`, that distinguishes between *Adjacent* instances (94 observations) and *Non-Adjacent* instances (395 observations), contrasting cases where the verb is adjacent to its referent, (5a, b), vs. all possible non-adjacent contexts (5c–l). As we shall see, assessing which of these best accounts for the variation in the data is a key to understanding the phenomenon.

Finally, we configure a predictor to test for the different types of determiner phrases in which the plural referent is contained, labelled as `DP Constituency` in the data file. This predictor tests for whether the internal structure of the `DP` has an effect on the realization of the verb. For example, Milsark (Milsark, 1977) proposed that strong determiners (i.e. definite articles) which undergo quantifier raising and adjoin to IP permit the verb to get agreement, i.e. *were* while weak determiners (i.e. numeric quantifiers, partitive constructions) do not. While this particular hypothesis was not borne out in the earlier analysis of this feature in York, some other divide within the range of different `DP` types may remain to be discovered. Thus, we differentiated the following configurations: a bare plural NP, (5a), those with a single modifying adjective, quantifier, partitive construction or combinations thereof (5b–i), *no* negation, (5d), and contexts with adverbs, (5h). In this categorization of the data the number of observations for the different factor levels ranges from 4 (for quantifiers functioning as pronouns), e.g. *There weren't many*, to 100 for partitive constructions (5g), e.g. *There was lots of cinemas.*[3]

The utterance *There was these two other lads* was coded as follows: `Proximate1` = 0; `Proximate2` = 3, `Prox1.adj` = *Adjacent*; `Adjacency` = *Non-Adjacent*, `DP Constituency` = *definite*. The different ways of encoding proximity give rise to a set of highly collinear (non-orthogonal) predictors. Furthermore, neither `Proximate1` nor `Proximate2` and DP constituency are fully independent: A large majority (68 out of 94) of the observations labeled as adjacent for the factor `Adjacency` are bare NPs (see Table 1). Such interdependencies between predictors is a common phenomenon in sociolinguistic data sets since competing hypotheses are inevitably co-dependent. We will discuss below what options there are to explore such co-dependent predictors from a statistical perspective.

DP Constituency	Adjacent	Non-Adjacent	Examples
adjective	0	21	5 b
all	0	8	5 g
bare adjective NP	0	1	
bare NP	68	2	5 a
combination	0	70	5 k
definite	0	29	5 d
negation	0	25	5 e
numeric quantifier	18	66	5 c
partitive	1	99	5 h
non-numeric quantifier	4	52	5 f
adverb	3	22	5m

Table 1: Contingency table for the predictors `DP Constituency` and `Adjacency`

4 Statistical modeling

It is uncontroversial that appropriate statistical tests should be conducted in the analysis of linguistic variation and change. Such tests enable the analyst to determine whether the patterns observed in the data are the product of chance or not. Sociolinguistics was the first subfield of linguistics to embrace the use of the generalized linear model, implemented in the Varbrul software of the 1970's (see, e.g. Sankoff and Sankoff, 1973; Cedergren and Sankoff, 1974; Rousseau and Sankoff, 1978; Sankoff, 1978c,b; Sankoff and Laberge, 1978; Sankoff and Labov, 1979; Sankoff and

[3]Discourse markers and/or pragmatic expressions were counted as intervening words but ignored in coding for DP constituency. Thus, contexts such as, e.g., *There was like, as I say, three of us.* (YRK/89), were coded as having four intervening words for `Proximate1`, six intervening words for `Proximate2` and partitive for `DP Constituency`.

Rousseau, 1979; Sankoff, 1982, 1985, 1978a). An early general software package for the generalized linear model was GLIM (Nelder, 1975). Since then, the generalized linear model has become available in any of the major statistical software packages, including SAS, SPSS, and R. In the present study, we consider two new tools in statistics that have reached maturity in the last decade: the generalized linear *mixed-effects* model and *random forests*. We believe both tools have valuable advantages to offer for the analysis of sociolinguistic data.

Sociolinguists find themselves confronted with many data related problems. Among these is the well known fact that the data are almost always more sparse than is desirable and are typically unevenly distributed across individuals, social groups and linguistic contexts. Moreover, the data always displays a great deal of variation, both inter-individual and intra-individual. Many data sets come with categorical individuals, i.e. those without any variation in the phenomenon of interest, and inevitably the data are characterized by many unfilled, empty cells and inevitably cells with just one observation (singletons). A case in point is the 1998 data on *was/were* variation. The subset of the data upon which we focus in this analysis — plural existentials — comprised only 310 tokens ranging from 1–29 tokens per individual (Tagliamonte, 1998).

A generalized linear model is ideal for handling many kinds of data sets. However, the new mixed-effects models provide the researcher with an even more powerful and principled way of dealing with different kinds of predictors typically encountered in sociolinguistic data sets. Consider the individual speakers or writers typically sampled in sociolinguistic studies. Such individuals are generally sampled from larger populations of similar individuals, and are selected to be representative for these larger populations (e.g., the city, region or dialect individuals come from). This brings us to a central distinction in the statistical analysis of factorial data (i.e. data that can be categorized into levels or factors). There is an essential difference between fixed-effect and random-effect factorial predictors. An example of a fixed-effect factor is the sex of the individual, which has exactly two levels (female versus male) and exhausts all possible levels for this predictor, at least from a biological perspective. Random-effect predictors, by contrast, have levels that constitute only a subset of the possible categories available in the population. Individuals (and also words, e.g., nouns, verbs or adjectives) are typical examples of random-effect factors. If, in a statistical analysis, a random-effect predictor is analysed as if it were a fixed-effect predictor, then the conclusions reached will only be valid for the individuals and words sampled. Thus, if the sample comprises 8 individuals the statistical model will be valid for only those 8 individuals. Conclusions do not automatically generalize to the relevant populations of interest. For generalization, p-values may be too small and misleading.

A random-effect factor such as Individual can be treated as fixed only when each individual contributes a single observation. In other words, for data sets that sample just one utterance from a given individual and that record only a single instance of a given utterance, the distinction between fixed and random factors is not at issue. In this case, the traditional generalized linear model is an excellent choice. Perhaps the best example of such a study is Labov's famous department store research where many people were asked a question that lead them to say 'fourth floor' (Labov, 1972b) so that variation in the pronunciation of [r] could be analyzed.

However, most sociolinguistic studies are not designed in this way. Instead, there are a limited number of individuals and (hopefully) many tokens from each individual. This presents a problem for statistical modelling because as soon as a given individual contributes more than one observation, the individual him or herself becomes a source of variation that should be brought into the statistical model. Consider a group of individuals from the same age group, with the same sex and the same education level. Within such a socially homogeneous group,it is possible the rate of a linguistic variant such as *was* will have widely diverging preferences across individuals.

If the individual is not considered as a predictor in the model and the individuals in the data use a variant with widely diverging individual probabilities, two things may go wrong. First, we may miss out on the opportunity to better understand the data, and to explain more of the variation. For example, a variable that remains constant across a population, i.e. no effect of individual variation, will require a different interpretation than a variable where the individuals exert so much of an effect than none of the other predictors are significant! Second, the data will have correlated errors (deviances). To see this, consider again our group of individuals that are homogeneous with respect to Age, Education, and Sex. All observations for the individual strongly favoring *was* will have large deviations from the group mean. Conversely, for an individual strongly disfavoring *was*, large deviations in the opposite direction will be present. This will undoubtedly arise when the use of a particular variant is socially restricted in some way. For example, the use of *ain't* or the words *supper* or *tea* for the evening meal. Some individuals may use them; others may not. Best practice in variationist methodology is to examine cross-tabulations of internal and external predictors for each individual in the data sample in order to evaluate what effect each individual may have (Tagliamonte, 2006). Indeed, there is a rich literature on the anomalous behaviours of individuals inside community-based samples (e.g. lames) (Labov, 1972a), oddballs (Chambers, 1998, p. 94)), and strategies have been proposed to find and evaluate the effect such individuals may have on the data (van de Velde and van Hout, 1998; Guy, 1980). The main point we are making here is that mixed-effects modeling allows the researcher to incorporate some of these insights straightforwardly into the statistical model. Although mixed-effects models can bring individual differences into the statistical model, they do not protect protection against distortion by atypical outliers. Model criticism is an essential part of good statistical practise, irrespective of whether a mixed-effects approach is adopted.

Figure 1 illustrates this problem for a classic generalized linear model (left panel) and compares this with a generalized linear mixed model (right panel) fitted to the *was/were* data. The basic idea here is that some individuals behave consistently (favoring or disfavoring a particular variant) in ways which cannot be explained as resulting from the combination of social characteristics they are coded for. We will discuss the details of these models below. Here, we draw attention to the distribution of the deviance residuals. The deviance residuals concern the difference between the observed and predicted values.

Each box and whiskers plot in Figure 1 represents the deviance residuals for a given individual, labeled `a` through `p`. The box represents the interquartile range (50% of the data points), the whiskers extend to 1.5 times the interquartile range, and circles represent individual outliers. Ideally, the boxes should be centered symmetrically around the zero line. For many individuals, this is the case, even though the medians (represented by solid black dots) tend to be relatively far away from zero, with subject `g` as the only exception. What is of special interest to us here are exceptions such as individual `b` (all her deviance residuals are negative) and individual `h` (all her deviance residuals are positive). Both these individuals are extreme, in that individual `b` always uses *were*, and individual `h` always uses *was*. A model ignoring this variation among the individuals fails to fit the data of the individuals accurately. As we shall see below, the model was informed about Adjacency, Polarity, and Age, but these predictors do not provide enough information about the preferences of individuals. Given the information it has, the model does not expect such a strong preference of individual `b` for *were*, nor such a strong preference of individual `h` for *was*. For individual `h`, it underestimates her preference for *was*, hence the positive residuals. For individual `b`, it underestimates her preference for *were*, hence the negative residuals.

It is only when the individual is brought into the model as an explanatory factor that it becomes possible to correct for this systematic prediction error of the standard logistic model.

The right panel of Figure 1 shows that the estimation errors of the corresponding mixed model are much reduced, thanks to the inclusion of individual as a (random-effect) predictor in the generalized linear mixed model: The median deviances are closer to zero, indicating that a better fit of the model to the data has been obtained. A mixed effects model with individual as a random effect offers the analyst a statistical validation of the significance of the social and linguistic factors in the model over and f the effect of individual. Using the standard logistic model might have led to the exclusion of extreme individuals in a sample; however, the mixed-effects model enables the analysis to proceed further.

Note that the mixed model has a few more extreme outliers for individuals k and l, but then most of the points (concentrated in or close to the black dots) are much closer to zero. In other words, a slight increase in deviance for a few points is offset by a slight decrease in deviance for lots of points. The latter is better.[4]

Nevertheless, in some cases even mixed-effects models can be challenged by the often highly unequal numbers of tokens involved for different combinations of predictors. Some stress for the mixed model is clearly visible in the right panel of Figure 1: In the ideal situation a model's underlying assumptions are appropriate for the data and the medians should all be close to zero. The divergences from zero indicate that some of the assumptions underlying the mixed-effects model are violated.

Furthermore, the kind of interactions that a (mixed-effects) generalized linear model can handle effectively may for some data sets be too restricted for the highly imbalanced cells typical of sociolinguistic data. As we shall see, this is where conditional inference trees and random forests provide a complementary technique that may provide insights that are sometimes difficult or impossible to obtain with the linear model.

4.1 A generalized linear model

Table 2 presents the results of a variable rule analysis. It is a standard generalized linear model with four predictors: `Polarity, Adjacency, Sex`, and age. `Adjacency` taps into the proximity complex through a binary factorial predictor with as levels *Adjacent* vs. *Non-Adjacent*, as discussed above. `AgeGroup` is configured with four levels: 20–30, 31–50, 51–70, and 70+. The response variable is the binary choice between *was* and *were*. The model seeks to predict which variant is used, and considers *was* as a 'success', and *were* as a 'failure'. In other words, percentages and probabilities are calculated from the perspective of *was*, and the model evaluates how often *was* was used compared to all instances of *was* and *were* jointly. We consider a model with main effects only, excluding interactions. In statistics, such a model is referred to as a *simple main effects* model.

Table 2 provides the following information: The predictors considered in the analysis (Factors), their levels (Levels), the number of tokens (Counts), the number of cases with *was* (Successes), the percentage of such cases (Perc), the corresponding probabilities as estimated by the model (Probs), and the factor weights (Weight). This kind of output is familiar variable rule analyses. Now, let us consider the results from a general statistical perspective.

The results in Table 2 are based on a series of decisions that jointly define a very specific instantiation of the generalized linear model. An key distinction that is crucial to understanding the reportage in Table 2 is that between unordered and ordered factors. *Unordered factors* are factors with factor levels that cannot be ordered along a scale of magnitude. Polarity, Adjacency, and Sex are unordered factors. By contrast, Age is an ordered factor as its levels, 20–30, 31-50,

[4]When there is no box for an individual it means that the interquartile range is restricted to a single value (so very little variation in values).

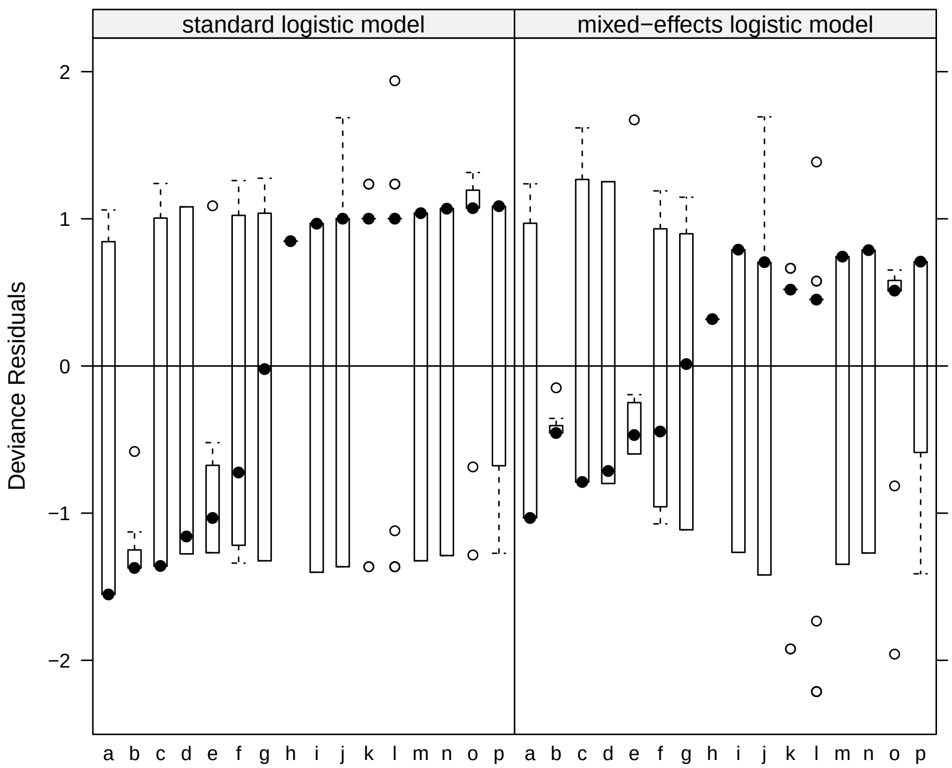

Figure 1: Deviance residuals for a standard logistic model (left) and a logistic mixed model (right) for individuals contributing at least 10 utterances. Each boxplot should be centered around zero (highlighted by the horizontal line).

	Factor	Levels	Successes	Counts	Perc	Probs	Weight
1	Polarity	Affirmative	270	455	59.34	0.5852	64.46
2	Polarity	Negative	10	34	29.41	0.3054	36.48
3	Adjacency	Adjacent	40	94	42.55	0.3655	42.49
4	Adjacency	Non-Adjacent	240	395	60.76	0.5185	57.79
5	Sex	F	161	270	59.63	0.4761	53.55
6	Sex	M	119	219	54.34	0.4057	46.51
7	AgeGroup	20-30	62	77	80.52	0.7061	70.61
8	AgeGroup	31-50	36	62	58.06	0.4827	48.27
9	AgeGroup	51-70	112	208	53.85	0.4208	42.08
10	AgeGroup	70+	70	142	49.30	0.3804	38.04

Table 2: Variable rule analysis, sum coding, all individuals, N=489; Counts: total number of observations.

51–70 and 70+ are on a scale from small (young) to large (old).

For unordered factors, the model uses what is known as *sum coding*. As a consequence, the factor weights (in the column Weight in Table 2) are differences from the grand mean, repositioned around 50%. For ordered factors, variable rule analysis implements polynomial contrasts. Polynomial contrasts are a good choice when the (ordered) predictor levels are equally spaced and have equal numbers of observations. The weights in Table 2 show decreasing probabilities for *was* with age.

Tables such as Table 2 do not report the coefficients of the underlying generalized linear model, which are on the logit (log odds) scale. This tradition makes a variable rule analysis stand out from the kind of statistical models generally reported in social science or science which may discourage cross-fertilization of knowledge across fields.[5]

What information does Table 2 reveal? First, the predictor `Polarity`, which has two levels, *Affirmative* and *Negative*, receives a weight greater than 50 for *Affirmative*, and a weight smaller than 50 for the level *Negative*. This indicates that according to the model the use of *was* is more likely for affirmative polarity, and less likely for negative polarity. This prediction of the model fits well with the observed counts of successes (uses of *was*) given the total numbers of observations. For affirmative polarity, 270 out of 455 observations have *was*, or 59.3%. For negative polarity, only 10/34 = 29% of the observations support *was*. The column labelled 'Probs' lists the proportions predicted by the model given the factor weights it estimated. It is easy to see that the predicted proportions are quite similar to the observed percentages. For the predictor `Adjacency`, the second unordered predictor in the model, the likelihood of *was* is slightly greater for non-adjacent contexts, and slightly smaller in adjacent contexts. For the predictor Sex, the factor weights are both close to 50. As we will observe below that this predictor does not reach significance. Finally, for the ordered factor `AgeGroup`, we see that as we move down the ordered predictor levels, from the youngest to the oldest group, the factor weights (and the observed percentages and predicted proportions) of *was* decrease.

Table 3 presents the results of an analysis using the same statistical tool, the generalized linear model for binary response variables, but now we apply it in a slightly different way. First,

[5]Using coefficients offers the analyst further possibilities for analysis. For example, the the coefficients estimated for ordered factors can be used to evaluate whether trends across factor levels are linear or curvilinear, as we shall see.

instead of examining the effects of predictors on the percentage scale, we consider effects on the log odds scale. On the percentage scale, the 50% mark is the pivotal value, with values greater than 50% indicating a trend in favor of the use of *was*, and values below 50% indicating a trend against *was* and in favor of its counterpart, *were*. On the log odds scale, the value of zero takes over this pivotal role. Positive values now indicate support for *was*, and negative values support for *were*. Second, instead of using sum coding, we now make use of *treatment coding*. With treatment coding, one predictor level is selected as the baseline, the so-called reference level. R, when not explicitly instructed otherwise, will select as reference level that predictor level that is the initial one in the alphabetically sorted list of factor levels. For `Polarity`, the reference level is *Affirmative*, as 'affirmative' precedes 'negative' in the alphabet. For `Adjacency`, the reference level is *Adjacent*, and for `Sex`, it is *F(emale)*. The reference level for `AgeGroup` is *20–30*. Given a reference level, treatment coding instructs the generalized linear model to estimate the differences between the other predictor levels of a given predictor, and that predictor's reference level. For `Polarity`, it will estimate the difference between the mean log odds for *negative* observations and the mean log odds for the *affirmative* observations. For `Sex`, it will calculate the difference between the mean for the males and the mean for the females.

Which kind of dummy coding is selected is to some extent a matter of personal preference. A first advantage of treatment coding is the coefficients it estimates are well interpretable for unbalanced datasets. For unbalanced designs, dummy coding with sum coding has as a consequence that the interpretation of the coefficients as differences from the group mean is only approximately correct. As a result, the factor weights as listed in Table 2, which are derived from these coefficients, are also approximations. A second advantage of treatment coding is that the coefficients remain transparent when interactions with other factors and with covariates are included in the model specification. We return to interactions below.

The widely varying tokens (number of observations) for the different cells defined by `Adjacency`, `Polarity` and `AgeGroup` make treatment coding a natural choice for our data. For instance, there are 77 observations for the youngest age group, and 62, 208, and 142 for the subsequent age groups. For such an unbalanced dataset, the coefficients of the model are much easier to interpret than the coefficients obtained with sum coding and with polynomial contrasts for ordered factors (see, e.g., Venables and Ripley, 2002).

	Estimate	Standard Error	Z	p
Intercept	1.0509	0.3774	2.7847	0.0054
Polarity=Negative	-1.1656	0.4011	-2.9062	0.0037
Adjacency=Non-Adjacent	0.6257	0.2409	2.5975	0.0094
Sex=Male	-0.2862	0.1951	-1.4671	0.1423
Age=31-50	-0.9457	0.3996	-2.3668	0.0179
Age=51-70	-1.1962	0.3253	-3.6774	0.0002
Age=70+	-1.3645	0.3371	-4.0479	0.0001

Table 3: Generalized linear model with only main effects, using treating coding, all individuals, N=489

How do we interpret tables such as Table 3? The first row of this table lists the intercept, which represents the reference levels of all factorial predictors in the model simultaneously. In other words, the estimate for the intercept is the mean log odds for the youngest age group (20–30), Affirmative Polarity, Adjacent Adjacency, and Females. This estimate is positive, indicating that for this combination of predictor levels, *was* is used more often than *were*. The second column presents the standard error associated with this estimate. The standard error is a measure of the uncertainty about the estimate. The larger this uncertainty, the less confidence should be placed in the estimate. The third column presents the Z-score, obtained by dividing the estimate by its standard error. This score follows a normal distribution, allowing us to calculate the probability (listed in the fourth column) of observing a more extreme value for the coefficient. Formally, this test asks whether the intercept is significantly different from zero, i.e., a 50-50, random use of the two forms. Since for the intercept, this probability is small, we can conclude that in all likelihood young females use *was* significantly more often than chance in Affirmative Adjacent contexts.

The next row in Table 3 shows that observations with negative polarity have a mean log odds that is -1.17 below that of the intercept. The standard error and its associated Z-value show that this difference reaches significance, $p = .0037$. The group mean that we can calculate from this, 1.0509-1.1656 = -0.1147, is calibrated for the young age group, for females, and for Adjacent Adjacency. When we only change polarity, we get the group mean for negative polarity, youngsters, adjacent adjacency, and females. The second row of the table tells us that the difference between these two group means is significant.

In the next line only `Adjacency` changes from *Adjacent* to *Non-Adjacent*, so we get young + women + positive + non-adjacent. This illustrates that young women in affirmative polarity use *was* less often in Non-Adjacent constructions than in Adjacent constructions. This contrast is also significant. The next predictor, `Sex`, comes with a contrast suggesting that males use *was* less frequently than do females. However, the large standard error and the low Z-value suggest that this small difference is not significant.

Finally, the effect of `AgeGroup`, a predictor with four levels, appears in the table with three contrasts: There are three predictor levels other than the reference level (the youngest age group), and each of these factor levels is contrasted with the reference level, producing three differences between group means. As the age group increases from 31–50 to 70+, the magnitude of the estimated contrast increases. From the last column, where the p values are listed we can observe that each of these three age groups uses *was* significantly less often than the youngest age group.

Variable rule analysis with sum coding and our re-analysis using treatment coding make exactly the same predictions, even though these predictions are arrived at in slightly different ways. Figure 2 illustrates this graphically. The left panels present the *partial effects* of the predictors `Adjacency`, `Polarity`, and `AgeGroup` in the model using treatment coding. The partial effect of a predictor is the effect of that predictor when all other predictors in the model are held constant, where possible at a typical value. For factorial predictors, it is useful to make the reference level the one that comprises the majority of observations. In this way, graphical representations of the data will represent the effects calibrated for the largest possible number of data points (which, with many cells, might be a small minority of all data points.) For numerical covariates, it makes sense to choose the median of that covariate as typical value.

The upper left panel shows the partial effect of adjacency for females in the 51–70 age group, for affirmative polarity. This plot was obtained using the plot facilities for logistic regression models in the `rms` package of Harrell (2001), which also adds confidence intervals around the group means. Now consider the lower left panel, which presents the partial effect for `AgeGroup`. The likelihood of *was* decreases as age increases. The pattern as shown in this panel is calibrated for affirmative polarity, non-adjacency, and females. For adjacent constructions, we know from the top left panel that the likelihood of *was* decreases. Hence, to make the bottom left panel precise for adjacent constructions, all four points have to be shifted down according to the amount of contrast between the adjacent and non-adjacent groups in the top left panel.[6] This illustrates what it is to plot a *partial* effect: The effect is calibrated for specific values of all other predictors in the model. If the value of one of the other predictors changes, the points for the predictor under consideration must be re-calibrated as well.

The right panels of Figure 2 present the partial effects for the variable rule model. There are no adjustments listed underneath each panel. This is because in this model all effects are positioned around the grand mean. In the upper right panel, the group means for the two levels of `Adjacency` show the same difference as in the upper left panel, but they are positioned differently. The group means in the left panel represent actual cells in the design, namely, adjacent and non-adjacent Adjacency for females in the 51–70 `AgeGroup` under affirmative polarity. Because the group means in the upper right panel are calibrated with respect to the grand mean, they do not represent any of the cells in the design. All the other differences between the levels of other factors are averaged out. In other words, the right panels summarize the general trends, the left panels present the same trends but position them specifically anchored with respect to specific cells in the design.[7]

Underlyingly, irrespective of which kind of dummy coding is used, contrasts are estimated on the logit scale. Because the transformation from logits to proportions is non-linear, the magnitude of a contrast on the proportions scale will vary between sum coding and treatment coding. This is illustrated in Table 4 for the case of the youngest age group producing adjacent sentences, comparing the effect of `Polarity` on the logit and back-transformed proportions scale (as in a Goldvarb analysis). The difference here is small, but depending on the data, it can be quite substantial. This nonlinearity also affects the confidence intervals for the group means, which on the logit scale are symmetrical around the mean, but, as can be seen in Figure 2 for negative polarity, can become noticably asymmetrical on the proportions scale.

Coding	Scale	Affirmative	Negative	Difference
Treatment coding	logit	1.0509	-0.1147	1.1656
Sum coding	logit	0.3440	-0.8216	1.1656
Treatment coding	proportion	0.7409	0.4714	0.2696
Sum coding	proportion	0.5852	0.3054	0.2797

Table 4: Contrasts on the logit scale are identical for sum and treatment coding, but after back-transforming to proportions, differences based on centered factor levels (sum coding) are larger. For treatment coding, the contrast in polarity is that for the youngest age group (20-30) and adjacent sentences.

The two coding systems have both advantages and disadvantages. For balanced datasets, sum coding and polynomial contrasts for ordered factors make it possible to present effects as adjustments from a grand mean, which fits well with the formulation of variable rules in Cedergren and Sankoff (1974), for instance. Unfortunately, for unbalanced data sets, the mathematical interpretation of the coefficients is less straightforward, although for actual practice the differences are probably benign.

[6]On the underlying log odds scale, this statement is precise and correct. Because the transformation from log odds to proportions is nonlinear, the effects of the main effects on the proportion scale are approximate when compared across the three panels.

[7]A critical difference between sum coding and treatment coding arises when effects are evaluated on the proportion scale. On the logit scale, differences between any two factor levels are identical irrespective of which factor coding system is used. However, when group means are not represented on the logit scale, but on the proportion scale, i.e., when the logits are transformed into proportions, the two coding systems yield slightly different results.

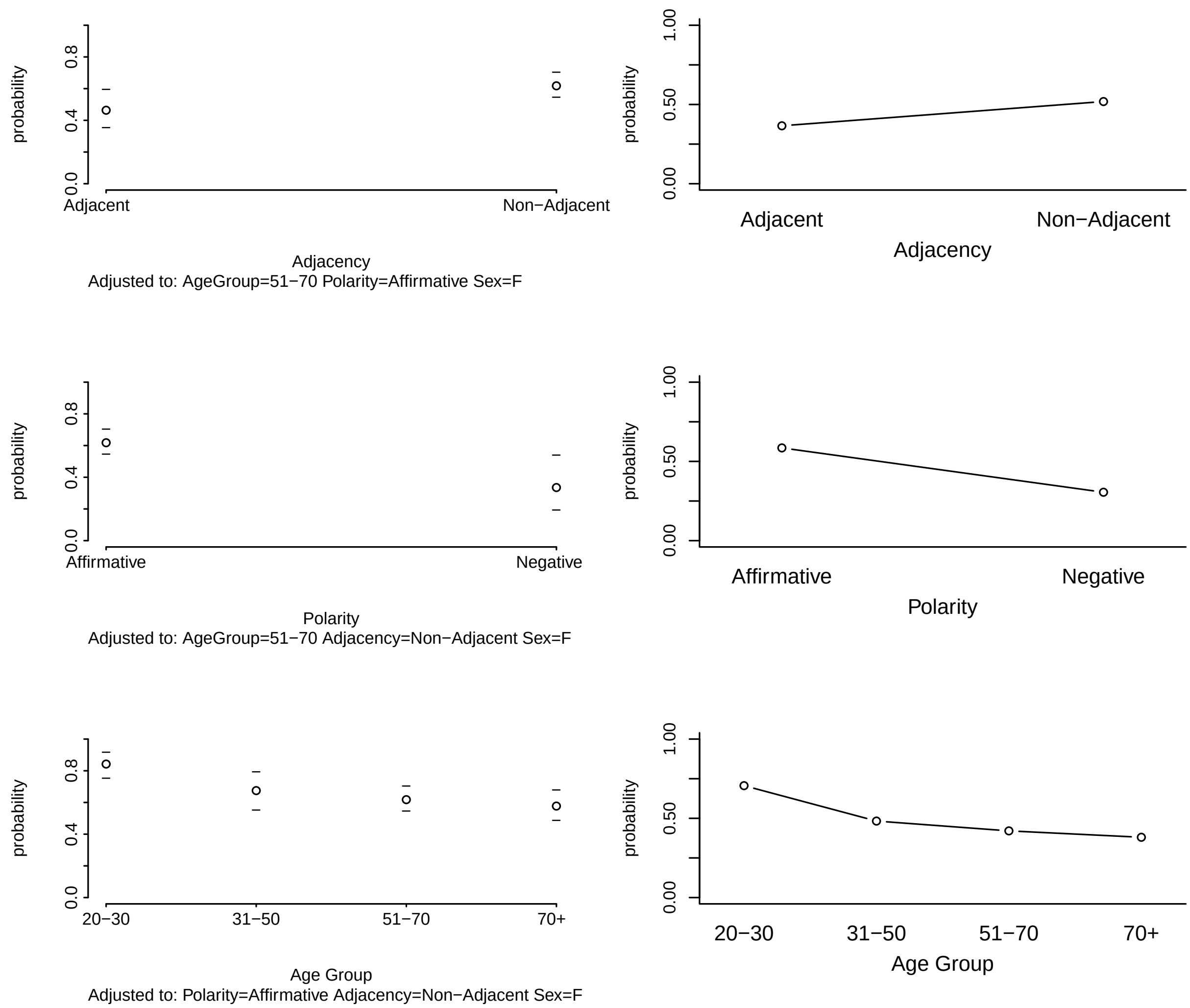

Figure 2: Effects of the predictors for a variable rule analysis with sum coding (right), and partial effects of the same predictors in a logistic model with treatment coding (left).

The advantage of treatment coding is that coefficients are well interpretable also for unbalanced designs, as often encountered when studying language. Furthermore, coefficients remain transparent when interactions and covariates are allowed into the model. The present data set, as with most sociolinguistic data sets, is in many ways highly unbalanced. As we shall see below, inclusion of both covariates and interactions in the model leads to improved prediction accuracy. Thus, we will use treatment coding for the remainder of this study.

4.2 Interactions and covariates

The model introduced in the preceding section (Tables 1 and 2) uses a factorial predictor, `AgeGroup`, to represent the age of the individuals, i.e. the predictor is divided into categories, a.k.a factors or levels. One disadvantage of this method is a potential loss of power, i.e., the likelihood of detecting an effect that is actually present in the data decreases (see, e.g. Baayen, 2010, and references cited there). Another disadvantage is that the cut-off points for the different age groups may be somewhat arbitrary, however carefully they may have been devised.

In Figure 2, the bottom panels indicate that the effect of `AgeGroup` is non-linear: The difference in the probabilities for the youngest age groups is larger than the corresponding difference for the oldest two age groups. This kind of trend, with an effect that becomes smaller with each step, is a negative decellerating trend. When replacing `AgeGroup` by `Age`, we should not expect to be able to model the negative decellerating effect of `Age` simply with a straight line

$$y = \beta_0 + \beta_1 x. \tag{1}$$

(In this equation, β_0 is the point on the vertical axis at which the line intersects the vertical axis, and β_1 is the slope of the line.) Instead of the formula for a straight line, what we need is a mathematical formula that can faithfully represent the observed curvature. In the present case, the curve looks like it might be part of a parabola. (In nature, the trajectory of a bouncing ball between two points where it touches the ground is part of a parabola.) Mathematically, a parabola is described by a quadratic polynomial, which adds a second parameter and a quadratic term to the equation of a straight line, as follows:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{2}$$

The coefficient β_2 is referred to as the *quadratic* coefficient as opposed to β_1, the *linear* coefficient.

Figure 3 illustrates the difference between a linear (solid line) and a quadratic fit (dotted line) for the trend across the four age groups visible in the lower panels of Figure 2. When modeling the effect of `AgeGroup` as a factor, no constraints are placed a priori on what kind of pattern the contrasts should follow. Some might be larger than the reference level, others smaller. When moving from an ordered factor (with imposed levels) to the underlying covariate, we may discover that the linear model is too restrictive. This can be seen in Figure 3, where the values on the horizontal axis range from 1 to 4, and the values on the vertical axis represent the proportions of *was* responses. The solid line represents a linear fit to the data, using for simplicity a standard Gaussian model. The dashed line represents a model with a quadratic polynomial. The amount of variance explained increases from 0.83 to 0.97. The addition of a second parameter allows us to model the observed trend more precisely.

Let us now test the difference between using a linear model for the factor `AgeGroup` and a quadratic model for the covariate `Age` (expressed in years), using logistic regression. Before we do so, recall that thus far, our statistical models have examined the data with only main effects. Such *simple main effects models* are correct only if each predictor has an effect that is independent of

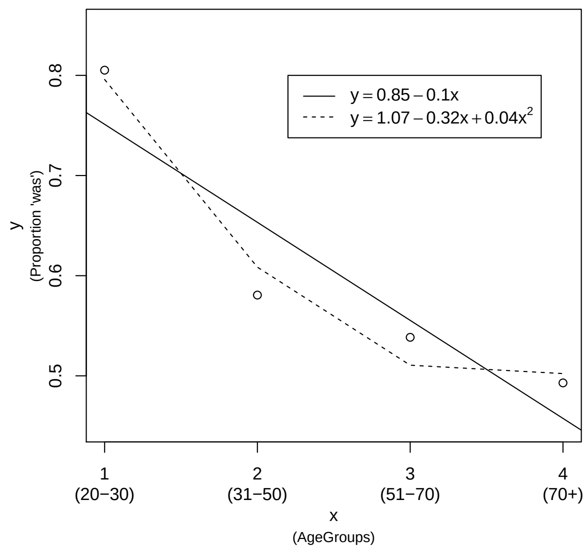

Figure 3: Linear and Quadratic fits to a non-linear trend (proportion *was* as a function of age group, compare the lower panels of Figure 2.

the effects of the other predictors. For any given data set, this assumption may or may not be appropriate. It turns out that the effect of `Age` differs for affirmative and negative polarity, and that a simple main effects model is too simple.

	Estimate	Standard Error	Z	p
Intercept	2.2779	0.7886	2.8887	0.0039
Adjacency=Non-Adjacent	0.6508	0.2412	2.6983	0.0070
Polarity=Negative	-17.3824	9.2231	-1.8847	0.0595
Age (linear)	-0.0793	0.0301	-2.6339	0.0084
Age (quadratic)	0.0006	0.0003	2.0916	0.0365
Polarity=Negative : Age (linear)	0.7171	0.3644	1.9677	0.0491
Polarity = Negative : Age (quadratic)	-0.0072	0.0035	-2.0609	0.0393

Table 5: Estimated coefficients, standard errors, Z and p values for a generalized linear model with an polynomial (degree 2) for Age in interaction with Polarity, using treatment coding, N=489.

Table 5 shows the coefficients of a model that includes linear and quadratic terms for Age, and that allows both these terms to interact with Polarity. (In this model, the predictor `Sex` is not included, because the preceding analyses indicated it does not reach significance.) The easiest way to understand what the model does is to inspect the visual representation of the interaction of `Age` by `Polarity` presented in Figure 4. The black lines represent the effect of `Age` for affirmative polarity and its 95% confidence intervals. As Age increases, the probability of *was* decreases. The gray lines represent the effect of `Age` for negative polarity. Since there are only 34 observations with negative polarity, compared to 455 observations with affirmative polarity, the confidence intervals are much wider, and no definite conclusions should be drawn from the analysis. However, the trend that we see in Figure 4 is that in utterances with negative polarity, *was* is favored by individuals around 50 years of age, and disfavored by the youngest and oldest individuals. This is a classic age-grading pattern and it shows us that affirmative and negative contexts reflect socially independent phenomena in this speech community.

Now consider the interpretation of the coefficients listed in Table 5. The intercept represents Adjacent Affirmative construction for individuals with Age zero. There are, of course, no such individuals in our sample. All data points are located far to the right of the vertical axis. Nevertheless, mathematically, the regression curves will intersect the vertical axis at some point, and for the Adjacent Affirmative constructions, this point is 2.28. The positive and significant contrast coefficient $(0.65, p = 0.007)$ for `Adjacency`=*Non-Adjacent* indicates that the probability of *was* increases for non-adjacent constructions compared to adjacent constructions (for age zero). The third row of the table indicates that for negative polarity, the likelihood of *was* decreases substantially $(-17.4, p = 0.06)$, again for age zero. (There are few data points here, so the standard error is large and the effect does not reach full significance.) [8]

The last two rows of Table 5 provide treatment contrasts that change the black curve in Figure 4 into the gray curve in Figure 4. On the logit scale, the black curve is given by

$$\text{log odds} = 2.2779 - 0.0793 \cdot \text{Age} + 0.0006 \cdot \text{Age}^2, \tag{3}$$

and the gray curve for negative polarity is given by

$$\text{log odds} = [2.2779 - 17.3824] + [-0.0793 + 0.7171] \cdot \text{Age} + [0.0006 - 0.0072] \cdot \text{Age}^2. \tag{4}$$

Note that all three coefficients in (3) are adjusted for negative polarity: the intercept, the linear coefficient of Age, and the quadratic coefficient of Age. When the parabola defined by (4) on the logit scale is transformed to the probability scale, the gray curve of Figure 4 results.

The model summarized in Table 5 invests no less that 5 parameters for the modeling of `Polarity` and `Age`. Does this investment pay off by leading to a model that fits the data better? This question can be answered in two ways. First, we can compare the present model with a much simpler model with simple main effects for `Adjacency, Polarity`, and `Age`. This model requires only four parameters: an intercept, two contrast coefficients, and one slope (see Table 6).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.0035	0.3777	2.6567	0.0079
AdjacencyNon-Adjacent	0.6551	0.2390	2.7407	0.0061
PolarityNegative	-1.1494	0.3964	-2.8993	0.0037
Age	-0.0197	0.0051	-3.8552	0.0001

Table 6: A main effects model with Adjacency, Polarity, and Age as predictors.

However, upon inspection it turns out that the residual deviance for this simpler model, 631.28, exceeds the residual deviance of the complex model, 616.76, by 14.52. This reduction in deviance follows a chi-squared distribution with as degrees of freedom the difference in the number of parameters, 3. The associated p-value, 0.002 obtained with this *analysis of deviance* indicates that the more complex model provides a significantly better goodness of fit. (For the example code in R, the reader is referred to the appendix.)

[8] For ages greater than zero, the linear and quadratic coefficients for Age (rows four and five) specify the parabola for affirmative polarity. They define the black curve in Figure 4. On the log-odds scale, this curve is part of a parabola. After transforming log-odds into the probabilities shown on the vertical axis, the curve remains U-shaped, but it is no longer a perfect parabola.

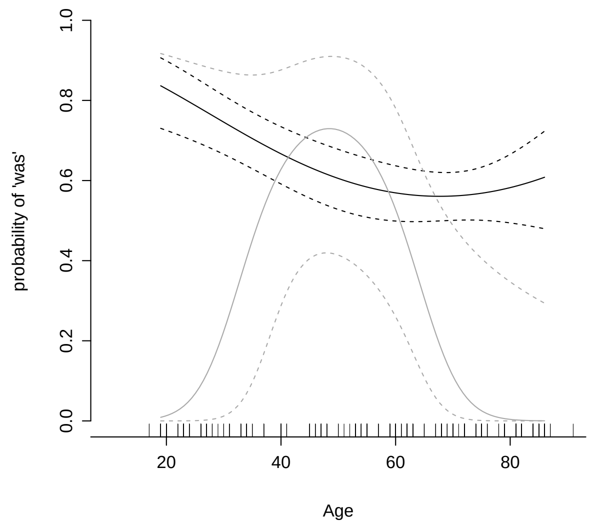

Figure 4: Partial effect of Age for sentences with affirmative (black) and negative (gray) polarity, with 95% confidence intervals.

Second, we can also compare the model of Table 5 with the original model with `Adjacency, Polarity, Sex`, and `AgeGroup` as predictors. That model also invested 7 coefficients (see Table 3). In this case, an analysis of deviance cannot be applied because both models invest the same number of parameters, and also because the models to be compared are not nested. For situations like this, it is often useful to use the index of concordance C. This index is a generalization of the area under the Receiver Operating Characteristic curve in signal detection theory (for examples, see, e.g., Harrell, 2001; Baayen, 2008). It measures how well the model discriminates between the *was* and *were* responses. When $C = 0.5$, classification performance is at change, values of $C \geq 0.8$ indicate a good performance. For the simple main effects model with `Adjacency, Polarity, Sex`, and `AgeGroup`, $C =$ 0.659. For the model of Table 5, there a slight improvement to $C =$ 0.66. For both models, however, the low value of C is a signal that the fit of the model to the data is not particularly good, which means that we have not yet arrived at a satisfying model to help interpret and explain the variation. One possibility is that there is simply a lot of noise in the data, and that this is the best we can do. Alternatively, it is possible that we are neglecting to enlist an alternative, and accessible, statistical tool, and that a much better fit is actually within reach.

4.3 Generalized linear mixed-effects modeling

In our analyses so far, we have not considered the individuals in the sample. These individuals will undoubtedly differ in their own personal preferences for *was* versus *were*. The question is, to what extent? Since the individuals contributing to the current data set are a small sample (83 individuals) of the locally born population of the city of York, `Individual` is a *random-effect* factor. Random-effect factors differ from *fixed-effect* factors such as `Adjacency` or `AgeGroup` in that the latter have a fixed and usually small number of factor levels that are *repeatable*. Repeatable in this sense contrasts `Adjacency`, which would have the very same factor levels in a replication study, namely *Adjacent* versus *Non-Adjacent*, with `Individual`, which in a new random sample would likely contain an entirely different set of individuals.

Variable rule analysis can only model `Individual` as a fixed effect in the existing data set. This has several disadvantages compared to mixed-effects models. First, it is not possible to generalize to the population that the data set is meant to represent. The model pertains only to the individuals who happened to be included in the sample.

Second, the estimated effects for the individual do not benefit from shrinkage. An individual evidencing an extreme preference for *were* in one random sample of elicited utterances is likely to show a reduced preference for *were* in a second random sample. This is an instance of the general phenomenon of regression towards the mean (often illustrated with the example that sons of very tall fathers tend to be less tall than their fathers). Shrinkage anticipates regression towards the mean, providing estimates for the individual differences that are more realistic, more precise, and hence afford enhanced prediction for replication studies with the same individuals.

Third, mixed models offer a flexible way of taking into account not only that individuals may have different preferences, but also that their sensitivity to, for instance, `Polarity`, may differ significantly. We return to this point in more detail below.

As a first mixed-effects model for our data, we begin with the model of Table 5 and simply add in `Individual` as a random-effect factor, allowing the intercept to be adjusted for each individual separately. This model with by-individual random intercepts assumes that the effects of `Adjacency` and `Polarity` are the same across all individuals, but allows individuals to have different baseline preferences for *was* versus *were*. Table 7 presents the coefficients for `Adjacency`, `Polarity`, and `Age` and their associated statistics.

What is new in the mixed model is an additional parameter that specifies how variable the individuals are with respect to their baseline preferences. This parameter, a standard deviation, was estimated at 1.33. This standard deviation squared is the variance of the individual baseline preferences.

	Estimate	Standard Error	Z	p
Intercept	-0.0712	0.3027	-0.2353	0.8140
Adjacency=Non-Adjacent	0.7389	0.2835	2.6066	0.0091
Polarity=Negative	-3.2308	1.2752	-2.5337	0.0113
Age (linear)	-10.7868	4.3684	-2.4693	0.0135
Age (quadratic)	8.3213	4.4844	1.8556	0.0635
Polarity=Negative : Age (linear)	-21.9671	20.0015	-1.0983	0.2721
Polarity = Negative : Age (quadratic)	-75.3279	32.2420	-2.3363	0.0195

Table 7: Coefficients of a generalized linear mixed-effects model with random intercepts for individuals (standard deviation 1.3345), using treatment coding and an orthogonal polynomial of degree 2 for Age.

Does the addition of this new parameter lead to an improved goodness of fit? This question is answered by comparing the original model (refitted with an orthogonal polynomial) with its mixed counterpart, using an analysis of deviance test. As the deviance is substantially reduced, from 616.76 to 565.02, it is not surprising that the mixed model provides a significantly better fit ($X^2_{(1)}$ =51.742, $p < 0.0001$). The index of concordance C increases substantially to 0.87, well above 0.8, providing statistical validation of a good fit. Finally, we have achieved an acceptable statistical model.

It is an empirical question whether by-individual random intercepts (or any further more complex random-effects structure) are justified for a given data set. When only a single observation is available per individual, it is not possible to include the individual as a random-effect factor. In that case, a generalized linear model suffices. However, for many practical situations, collecting only a single instance from each individual is prohibitively costly. Although variationist practice sometimes advocates restricting the number tokens per type for each individual Wolfram (1993), from a technical perspective, even a small number of by-individual replications causes problems for the classical statistical model. An advantage of the mixed-effects modeling framework is that it allows the researcher to sample as many tokens from a given individual as is feasible, thereby increasing statistical power. This also opens up additional possibilities to study how individuals differ systematically over and above the differences between the groups to which they belong. This may turn out to be useful for understanding the variation at the speech community level.

Returning to our data, it is worth noting that to this point we have assumed that the only difference between the individuals is their baseline preference for *was* versus *were*. However, there is some indication of significant variability in the sensitivity of the individuals to `Polarity`, which emerged in Figure 4 as linked to individuals' age, and which in our current mixed-effects model is accounted for by an interaction of `Polarity` by `Age`. When we relax the assumption that the effect of polarity is exactly the same for all individuals by allowing by-individual random contrasts for `Polarity` into the model specification, we obtain a model with a significantly improved goodness of fit, according to a likelihood ratio test ($X^2_{(2)} = 7.91, p = 0.0191$). Nevertheless, we are skating on thin ice. More than half of the individuals do not have a single negative token. The remaining individuals typically provide only a single example, with a maximum of four. Unfortunately, the paucity of data does not warrant exploring individual differences in their grammars for `Polarity`.

4.4 Random forests

Consequently, we turn to a relatively new tool: *random forests.* Random forests were developed by Breiman (2001), building on earlier work on classification and regression trees (Breiman, Friedman, Olshen, and Stone, 1984). In what follows, we make use of the implementation of random forests available in the party package in R (Strobl, Boulesteix, Kneib, Augustin, and Zeileis, 2008; Strobl, Boulesteix, Zeileis, and Hothorn, 2007; Hothorn, Buehlmann, Dudoit, Molinaro, and Van Der Laan, 2006a), which implements forests of conditional inference trees (Hothorn, Hornik, and Zeileis, 2006b). Like logistic models, random forests seek to predict, given a set of predictors, which of the alternatives *was* and *were* is most probable. However, these statistical techniques achieve the same goal quite differently. Logistic models predict the choice between *was* and *were* on the basis of a mathematical equation such as (3) above which specifies for each predictor how it affects this choice. Thanks to various simplifying assumptions, the mathematics of these models make it possible to estimate the parameters quickly and efficiently.

Random forests, in contrast, work through the data and, by trial and error, establish whether a variable is a useful predictor. The basic algorithm used by the random forests constructs conditional inference trees. A conditional inference tree provides estimates of the likelihood of the value of the response variable (*was/were*) on the basis of a series of binary questions about the values of predictor variables. For instance, for `Adjacency`, it considers whether splitting the data into adjacent and non-adjacent utterances goes hand in hand with the creation of one set of data points where *was* is used more often, and another set where *were* is used more often. The algorithm works through all predictors, splitting (partitioning) the data into subsets where justified, and then recursively considers each of the subsets, until further splitting is not justified. In this way, the algorithm partitions the input space into subsets that are increasingly homogeneous with respect to the levels of the response variable.

The result of this recursive binary splitting of the data is a conditional inference tree. At any step of the recursive process of building such a tree, for each predictor, a test of independence of that predictor and the response is carried out. If the test indicates independence, then that predictor is useless for predicting the use of *was* or *were*. If the null hypothesis of independence is rejected, the predictor is apparently useful. If there are no useful predictors, the algorithm stops. If there is more than one useful predictor, the predictor with the strongest association with the response is selected, the p-value of the corresponding test is recorded, and a binary split on the basis of that variable is implemented. Conditional inference trees implement safeguards ensuring that the selection of relevant effects (predictors, variables) is not biased in favor of those with many levels (multiple factors in a factor group), or biased in favor of numeric predictors (e.g. age of the individuals).

Random forests construct a large number of conditional inference trees (the random forest). Each tree in the forest is grown for a subset of the data generated by randomly sampling without replacement (subsampling) from observations and predictors. The metaphor used in statistics is of putting part of the observed data into a bag. The data in the bag is referred to as the 'in-bag' observations. The data points that were not sampled are referred to as the 'out-of-bag' observations. The consequence of this procedure is that for each tree a training set (the in-bag observations) is paired with a test set (the out-of-bag observations). The accuracy of a tree's predictions is evaluated by comparing its predictions for the out-of-bag observations with the actual values observed for the out-of-bag observations.

To evaluate how useful a predictor is, a permutation variable importance measure is used. Suppose that a given predictor is associated with the response variable. For example, *were* (as opposed to *was*) is associated with adjacency. By randomly permuting the values of the predictor, the association with the response variable is broken. An artificial example illustrating this point is given in Table 8. For the observed adjacencies, all but one non-adjacent utterance is paired with *was*, and all adjacent utterances are paired with *were*. When the levels of Adjacency are randomly permuted, this difference between *was* and *were* is erased. In this example, after permutation, adjacent utterances occur equally often with both forms, and the same holds for the non-adjacent utterances.

Response	observed Adjacency	permuted Adjacency
was	non-adjacent	adjacent
were	adjacent	adjacent
were	adjacent	non-adjacent
was	non-adjacent	non-adjacent
was	non-adjacent	adjacent
were	adjacent	non-adjacent
were	non-adjacent	adjacent
was	non-adjacent	non-adjacent
was	non-adjacent	non-adjacent
were	adjacent	non-adjacent

Table 8: Example of how permuting the levels of a predictor can break its association with the response variable.

In random forests, the permuted predictor, together with all the other predictors, is used to predict the response for the out-of-bag observations. If the original, unpermuted predictor was truly associated with the response, i.e., if the original predictor is a significant predictor of the response, then a model with the permuted version of the original predictor must be a less accurate classifier. In other words, classification accuracy will decrease substantially if the original, unpermuted predictor was truly associated with the response. The extent to which the model becomes worse is a measure of the importance of a predictor. If the model hardly becomes worse, then a predictor is not relevant. However, if the model's performance decreases dramatically, we know that we have a vital predictor. Breiman (2001) therefore proposes the difference in prediction accuracy before and after permuting the predictor, averaged over all trees, as a measure for variable importance.

In the present study, we make use of an improvement of this measure, the conditional variable importance measure implemented in the `cforest` function of the `party` package. Strobl et al. (2008) show that Breiman's original permutation importance severely overestimates the importance of correlated predictor variables. They propose a conditional permutation scheme that protects the evaluation of a variable's importance against inflation. For instance, in the present study of *was/were* variation, `Age` is a sensible predictor. A variable such as income, which correlates with age (older people tend to have higher incomes) is not a sensible predictor. Without appropriate measures, a random forest would nevertheless assign income a high variable importance, whereas a simple linear model would immediately detect that income is irrelevant once age is incorporated as a predictor. The conditional permutation variable importance implemented in the `cforest` function of the `party` package correctly reports spurious predictors such as income to have a very low variable importance.

Having outlined how the importance of variables is gauged by random forests, we finally need to introduce how a random forest is used to obtain predictions. After all, we are now dealing not with a single tree, but with a forest of trees. The solution adopted by the random forest technology is to make use of a voting scheme. All the trees in the forest contribute a vote based on what each tree thinks is the most likely response outcome, *was* or *were*. The prediction of the tree is the outcome that receives the greatest proportion of the votes.

Random forests provide a useful complement to logistic modeling in three ways. First, because random forests work with samples of the predictors, they are especially well applicable to problems with more variables than observations, i.e. "small n large p" problems. This situation is the typical case in sociolinguistic research where many studies are based on a relatively small number of tokens (observations) and a large number of predictors. Second, subsampling combined with conditional permutation variable importance estimation reduces substantially the problem of collinearity (correlated factors) that can severely destabilize regression models (Belsley et al., 1980). Third, empty cells, linear constraints in the predictors, or perfect separation of response classes in particular combinations of predictors may render regression modeling, or the exploration of interactions in a regression model, impossible. Random forests do not have these estimation problems, making them the ideal panacea for the thorniest problems of variation analysis. The added value is that random forests allow the researcher to explore more aspects of the data and by consequence more insights into the explanation for variable processes. (For an excellent introduction to random forests, see Strobl, Malley, and Tutz (2009).)

A random forest for our data with just the four predictors `Adjacency`, `Polarity`, `Age`, and `Individual`, comes with an index of concordance for this model, $C = 0.88$ that already presents a slight improvement on the value (0.87) obtained for the corresponding mixed model summarized in Table 7. However, the real power of the random forest becomes apparent when we consider other predictors that are available, but that were not included in the analyses with the generalized linear model (Tables 1 and 2) due to covariation with other predictors, highly unequal cell counts, empty cells, etc. Figure 5 presents the variable importance for the predictors `Individual` (the

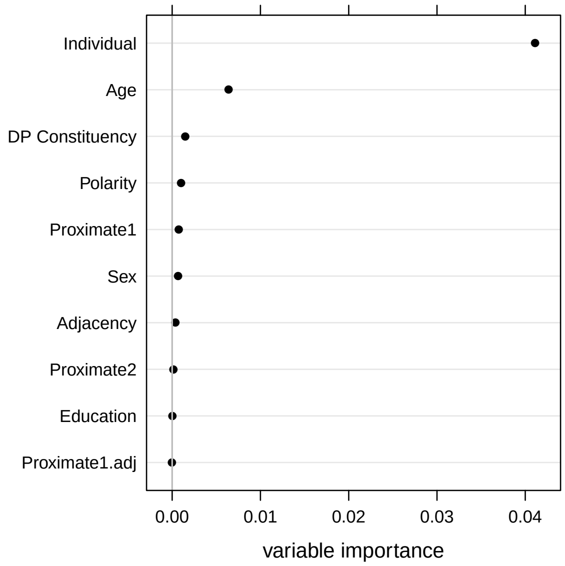

Figure 5: Conditional permutation variable importance for the random forest with all predictors. Predictors to the right of the rightmost vertical gray line are significant.

people in the sample), `Age` (the actual age of each person), `Polarity`, `DP Constituency` (11 levels, including levels such as *Bare NP, Numeric Quantifier, Partitive*, and *Definite*), the individuals' level of `Education` (*high* versus *low*), the `Sex` of the individual (*male* versus *female*), and the four different schemas for categorizing adjacency (described earlier). Note that within the linear modeling framework (including variable rule analysis), it would be impossible to explore simultaneously these highly correlated measures for `Proximity` and `DP constituency`. The index of concordance for the model with the full set of predictors increases to $C = 0.92$.

Figure 5 depicts the relative importance of the predictors, using conditional permutation-based variable importance. The gray vertical line highlights the variable importance of the inconsequential predictors, which is for all practical purposes equal to zero.

What Figure 5 shows is that the individual is by far the most important predictor. Substantial variability tied to the individual is also found in almost any psycholinguistic experiment (see, e.g. Baayen, 2008), where a subject random-effect factor invariably accounts for much of the variance. A key advantage of using mixed effects models for sociolinguistic studies will be the ability to amass a similar foundation of research. Analysts will be able to document the extent and nature of individual variance for linguistic features at all levels of grammar and across speech communities.

The next most-important predictor is `Age`, an external predictor also tied to the individual. Some predictivity is detectable for `DP constituency`, `Polarity`, `Proximate1`, and `Sex`. None of the other predictors contribute statistically significant effects, as indicated by the vertical gray line.

Before exploring how the predictor variables work together in predicting the choice between *was* and *were*, we emphasize again that the predictors considered jointly in this random forest are non-orthogonal and collinear. In particular, `Proximate1`, `Proximate2`, `Prox1.adj` and `Adjacency`, while not tapping into precisely the same underlying mechanism, are nonetheless highly collinear phenomena. Moreover, `DP Constituency` mirrors `Proximity` to a high degree since certain modifying structures in the DP are more complex and inevitably longer than others (e.g., quantifier phrase vs. bare NP). In a linear model, these predictors should never be considered together (see, e.g. Guy, 1988). Even when considered jointly in a (mixed) linear model, unsolvable computational problems arise, and error messages of various kinds are generated. The random forest, however, is immune to this kind of problem. It will consider all variables in their own right (factorial or numeric) and identify which of these variables is the superior predictor.[9]

Another useful property of the random forest is that it does not suffer from being excessively complex with too many predictors relative to number of observations (i.e. overfitting) and it is not hampered by small or even zero cell counts. For the present set of tests for proximity, `Proximate1` and `DP Constituency` are among the top three of the internal predictors in the analysis, together with `Sex`. These are among the most fine-grained predictors, one measuring distance in words to the plural referent and the other measuring the nature of the composition of the DP. Their relative importance reveals that the nature of the DP is a more germane predictor.

Thus, it becomes critical to understand the difference between these two predictors. `Proximate1` measures the proximity to a plural element and *was* is more likely in these contexts. `DP Constituency` identifies the different types of determiner phrases in the data. One of the most prominent types is partitive constructions (and combinations thereof), which are more likely to occur with *was* as well. Indeed, previous research has suggested a universal hierarchy of `DP Constituency`. So far, however, the rankings of categories have differed across studies (e.g., Hay and Schreier, 2004; Walker, 2007). This may be due to the fact that the distribution of DP types varies by data set or it may be due to varying coding strategies, but the fact that it turns up across studies is suggestive and in most cases the highest ranked category involves numbers *There was three of us; there was about fifty of us.* However, such constructions may or may not be grammatically plural despite the evident plural element. The relative ranking of `DP Constituency` in our analysis suggests that another underlying reason for variant *was* could be explained by certain NP constructions, in this case ones that are being reanalyzed as singular, not plural, hence *was* not *were*. In order to fully substantiate this hypothesis a more detailed semantic-syntactic analysis of `DP Constituency` is required.[10]

In order to clarify how the predictors evaluated by the random forest work together, we now consider the conditional inference tree for the data, grown with all predictors included. The superiority of a random forest (Figure 5) compared to a single conditional inference tree, grown with all predictors and all datapoints (see Figure 6) is evident from the inferior index of concordance for the single tree, $C = 0.79$. Nevertheless, the conditional inference tree highlights the complex interaction characterizing this data set: `Polarity` is relevant only for a subset of the individuals, and the effect of `Age` is further restricted to positive polarity items for that subset of individuals, in congruence with the linear model's evaluation of this interaction (cf. Figure 4).

Complex interactions, such as the one observed here, involving `Individual`, `Age`, and `Polarity`, can be difficult or even impossible to capture adequately even with a mixed-effects logistic linear model. In order to capture the differences between the individuals emerging from the conditional inference tree, the random-effects structure of the mixed-effects model would have to be enriched with by-individual random effects for `Polarity` and `Age`, as well as individual differences

[9] There is a vital difference in modeling an unordered (factorial) vs. an ordered (numeric) predictor. In the former the classification tree will try all possible splits of the data and there will be many different subsets. With a numeric predictor however, the model is much more constrained, due the intrinsic order of the factor levels. This means that the result of the analysis will be more linguistically sensible if the predictor is indeed well-characterized as ordered.

[10] The essential idea is that numerals as nouns are singular but as quantifiers they pluralize the noun. In other words, they have variable behaviour. In the case of partitive structures, it seems there is ambiguity about whether they involve multiple DPs or just one, with a quantifier, and this may be the reason for the current results (Massam p.c. 2.23.12).

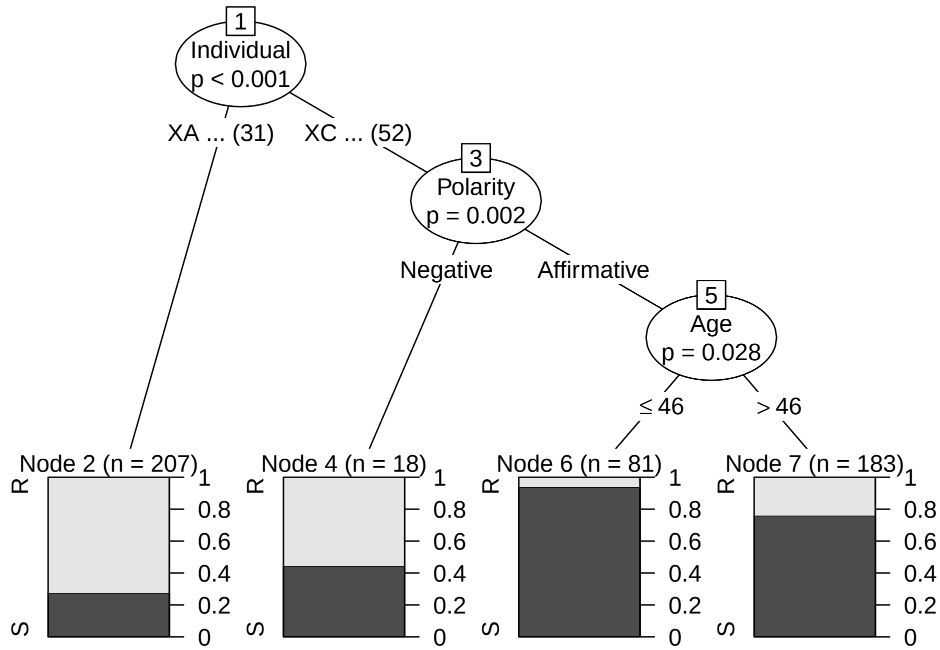

Figure 6: Conditional inference recursive partitioning tree.

for the interaction of `Polarity` by `Age`. Above, we have briefly mentioned that including random contrasts for `Polarity` improved the fit of the mixed model. But we also observed that there were very few examples of negative polarity in the data, which is why we did not pursue a more complex random effects structure. The conditional inference tree indicates that a much more complex random effects structure is required than we anticipated there. However, due to data sparsity, the mixed-effects model that we fitted to the data, with no less than 10 random effects parameters, was clearly stretched beyond its limits, and is not discussed further here. In contrast, the random forest and conditional inference tree offer an ideal tool to be used in tandem with the mixed-effects logistic model to come to a full understanding of the quantitative structure of a data set and as a result an optimal interpretation of the variation. In this case, we are pointed to the fine-grained distinctions among the predictors, particularly, `Proxmate1` and `DP Constituency`. Their relative importance points to the predictor that offers the better explanation for *was/were* variation and to which we should turn to inform our interpretation of the data.

In summary, for naturalistic, unbalanced data with complex interactions, random forests help overcome the limitations of mixed-effects models, although the reader should be warned that this comes at the cost of substantially more computing time. The smart mathematics underlying the mixed model make it possible to fit a model to the present data set in a few seconds. By contrast, even with smart computational optimization, the calculation of variable importance, based as it is on extensive permutation schemes, can take many hours to complete.

4.5 Restricting the analysis to variable individuals

The final question that we consider here is whether only variable individuals should be included in the analysis. In the present data set, there are 38 individuals who show no variation in their choice

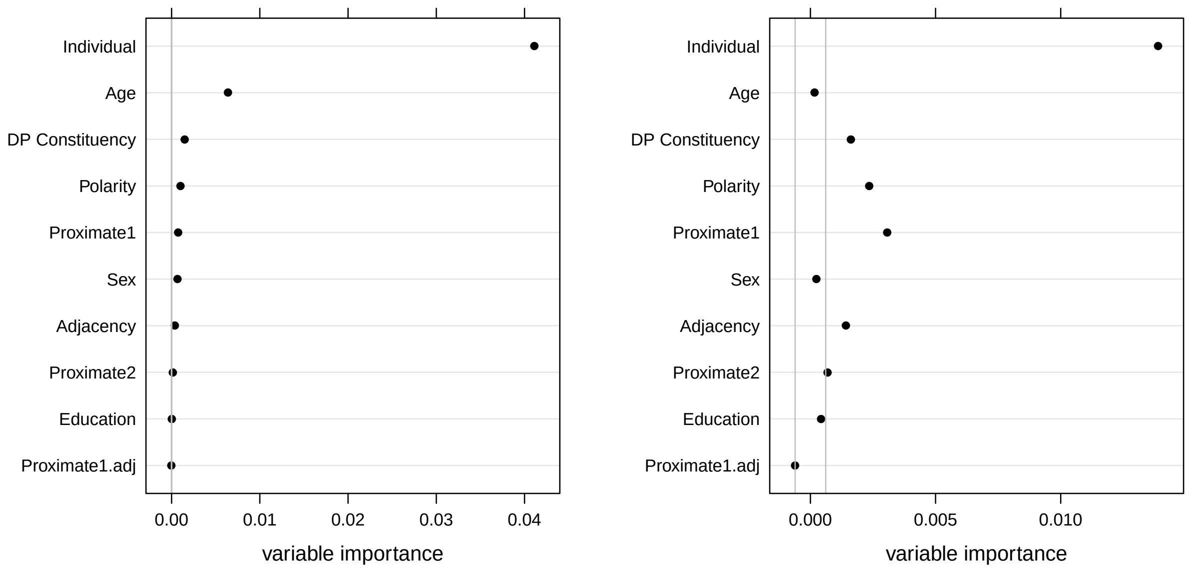

Figure 7: Conditional inference recursive partitioning trees for all individuals (left) and for variable individuals (right).

of *was* versus *were*. Variationist methodology typically recommends that categorical individuals be removed for the study of variable phenomena (e.g., Guy, 1988, p. 130). However, in practice, particularly with morpho-syntactic and discourse-pragmatic features, they are often included on the assumption that internal predictors will be parallel across individuals. The question is whether or not these individuals without variation are a source of noise that should be taken out before the start of the analysis? Would the relative importance of the predictors change if a random forest were fitted to the data after exclusion of the non-variable individuals?

	Factor	Levels	Successes	Counts	Perc	Probs	Weight
1	Polarity	Affirmative	195	326	59.82	0.5492	65.14
2	Polarity	Negative	10	31	32.26	0.2638	36.59
3	Adjacency	Adjacent	34	78	43.59	0.3171	41.92
4	Adjacency	Non-Adjacent	171	279	61.29	0.4846	58.68
5	Sex	F	119	191	62.30	0.4587	56.08
6	Sex	M	86	166	51.81	0.3401	44.22

Table 9: Variable rule analysis, sum coding, variable individuals only, N=357.

Table 9 shows a variable rule-style simple main effects model for the variable individuals only (compare Table 1; C =0.622). `AgeGroup` is not significant (and was therefore removed from the model specification). Instead, `Sex` now takes over as a significant external predictor. In a mixed-effects model including random intercepts for `Individual`, the effect of `Sex` is marginal ($p = 0.0621$, two-tailed test), but females favoured *was*. This is the result for `Sex` reported in the original study of *was/were* variation in York for a smaller set of individuals (Tagliamonte, 1998, p.181). When a random forest is grown for this subset of the data, the index of concordance C equals 0.88, a value that is lower than that for the random forest for all individuals ($C = 0.92$), but higher than the value reached by the model for all individuals when its predictions are evaluated for just the subset of data with variable individuals ($C = 0.78$). As can be seen in Figure 7, the importance of the variables changes as well. `Age` is now irrelevant, whereas `Polarity` and `Proximate1`, and to a lesser degree `Adjacency` and `DP Constituency` have gained importance.

In the right panel of Figure 7, two vertical gray lines are displayed. These lines have been added by hand to highlight the relative importance of the predictors. Those on the left line or below can be considered superfluous while those on the right are taken to be acceptable.

These changes indicate that the non-variable individuals are not just random noise. Being a non-variable individual must be, at least in part, predictable from the other variables. To pursue this possibility, we fitted both a conditional inference tree and a logistic model to the data with as a dependent variable whether the individuals did not show any variability (models not shown). The generalized linear model pointed to a highly significant effect of `Age` (older individuals are more variable, $p < 0.0001$) and possibly effects of `Polarity` (negative polarity increases variability, $p = 0.0446$) and `Adjacency` (non-adjacency decreases variability, $p = 0.0573$). With an index of concordance $C = 0.68$, this model did not outperform a conditional inference tree with a single split in age at 60: $C = 0.69$, see Table 10.

	age > 60	age ≤ 60
non-deterministic individual	252	105
deterministic individual	42	90

Table 10: Deterministic and variable individuals cross-classified by an age cut-off at 60, as suggested by a conditional inference tree.

This example illustrates the more general methodological point, namely, that the effect of categorial and non-categorical individuals should be brought into the analytical exploratory maneuvers of a variationist analysis (Guy, 1980). Are the categorical individuals random or can they be predicted by other variables? It makes sense to zoom in on what might be going on with variable informants otherwise valuable information might be swamped by noise. Now that statistical techniques are available which can easily include the individual as part of the analysis, sociolinguists will be able to deepen their understanding of the dialectic between individual and group behaviour.

5 Discussion

The use of plural existential *was* is a pervasive, highly variable, feature of contemporary varieties of English. The case study of York we have conducted here provides insight from a single speech community in a geographical setting - one of the oldest cities in northern England — where English has evolved in situ for centuries. In a 1998 study of *was/were* variation in this community, the analysis suggested two explanatory predictors — `Polarity` and `Adjacency`. Despite the unambiguous social values assigned to the variants, namely non-standard *was* and standard *were*, little explanatory value could be attributed to factors that typically provide a good measure of social embedding for linguistic variables of this type (e.g. `Education`, `Sex`). Instead, the results suggested that young females were leading an ongoing rise of existential *was*. However, the original analysis was based on only 310 tokens from 40 individuals and a fixed effects analysis.

The present analyses are based on an augmented data set, 489 tokens from 83 individuals, which at the outset provides for a better statistical model. By employing several new statistical tools we have gained an enriched view of this data. A mixed effects model enabled us to include `Individual` as a random effect factor and `Age` as a nonlinear numeric predictor with both linear and quadratic terms. This bolstered the original finding that two internal constraints — `Polarity` and `Adjacency` — underlie the realization of forms and that there is a bona fide change in progress. However, we have discovered a far greater source of explanation underlies the predictor labelled "Adjacency" than previously thought. In the random forest and conditional inference tree analyses we were able to model predictors that are continuous. A case in point is the relationship between the verb and the plural referent vs. its proximity to the closest plural element. When these were treated as independent continuous predictors (rather than factorial predictors) we discovered that; 1) they were more explanatory than any binary categorization of proximity, i.e. adjacent/non-adjacent (either as adjacent to the referent or the closest plural element); and 2) the relative importance of the DP complex, `DP Constituency` over proximity to a plural element, `Proximate1` was revealed. Finally, critical inter-relationships among social and linguistic factors have come to the fore, enabling new explanatory insights into the *was/were* variation in York and perhaps more generally, as we detail below.

A simple main effects model presented in Tables 2 and 3 was the starting point of our analyses; however, the index of concordance was only modest at 0.66. At the outset of our foray into new statistical tools, we first noted the difference between sum coding and treatment coding in presenting statistical results. Both kinds of dummy coding lead to the same predictions. They differ in that the former calibrates group differences with respect to the grand mean and the latter with respect to a 'default', the reference level. We used treatment coding as it offers more straightforwardly interpretable coefficients for understanding interactions between factors and covariates. We also moved towards testing the actual `Age` of each individual rather than working with a factor `AgeGroup`.

In exploring interactions in the data set using treatment coding and `Age` as a numeric covariate, we discovered a strong interaction between `Age` and `Polarity` (dramatically portrayed in Figure 4). While existential *was* was increasing monotonically in apparent time for affirmative contexts, confirming the earlier results, it is an inverse U-shaped curve for negative contexts, exposing a higher likelihood of *was* around 50 years of age. This only became evident when the analysis was expanded to include linear and quadratic terms for `Age` and their interaction with with `Polarity`, and yet there was only a tiny improvement, C =0.66, in how well the model discriminated between the *was* and *were* responses.

We then made the transition from a standard logistic model to a mixed-effects logistic model (Table 7) and included the individual as a random-effect factor. This tool offered several advantages. First, we obtained a much better fit of the model to the data, C =0.87. Second, including the individuals as a random-effect factor permitted us to be more confident about making generalizations from the data set at hand to the population it represents. Third, the mixed-model provided enhanced estimates of the coefficients and generally reduced standard errors for these estimates, resulting in smaller p-values and hence greater confidence that these coefficients are the most useful for formulating interpretations. Here, it is evident that the enhanced toolkit offers more solid statistical support for assisting interpretation of the data.

When we brought the individual into the model, we did this by allowing for adjustments to the intercept for the individuals. In this way, we could do justice to the slightly different baseline rates of *was* compared to *were* for individuals. We explored whether there might be additional differences between individuals and discovered that the effect of `Polarity` was highly circumscribed. Negative tokens of *was* are restricted to several of the uneducated women in the data. Due to this fact and the general scarcity of negatives in the data base, N = 34, we could not pursue individual differences further.

We complemented the mixed-effects logistic model with an analysis using random forests, a computationally intensive but high-precision non-parametric classifier. Fitting the same set of predictors to the data improved the index of concordance to 0.88. However, the real power of the random forest technique became apparent when we considered the full set of predictors that had been coded into the data files. The index of concordance rose to 0.92. Inspection of the importance of the predictor variables (Figure 5) bolstered the building evidence that `Individual`, `Age`, `Polarity`, `DP constituency` and `Proximate1` are the key factors in the realization of *was*. The novel contribution here is the nuanced perspective of the relative importance of all the potential predictors simultaneously.

An eminently useful property of random forests is that many different variables, even those that seek to capture similar underlying phenomena but use different factor levels (configurations), can be included and explored together. This is something that is not possible in logistic models. The models we have employed test several configurations that probe for proximity effects: (`Proximate1`, `Proximate2`, `Prox1.adj`, `Adjacency`). The binary predictors turned out not to be relevant and so did the proximity in words between the verb and its referent. Instead, Figure 5 and 7 show that `Proximate1` (the number in words intervening to the closest plural element) offers the most important contribution of this set of predictors (Figures 5 and 7). However, vying for importance is `DP Constituency` which exposed an underlying syntactic explanation.

The results arising from our analyses of variable and non-variable individuals which shows that the `Adjacency` predictor changes over the course of the current generation of speakers supports the idea that some kind of reanalysis of the number interpretation of the DP complex may be underway. All our analyses point to `Age` (as continuous) and `DP Constituency` as significant predictors. While it is beyond the scope of the present paper to conduct an in-depth syntactic analysis of the structure of each DP type, it suggests an interesting way forward for future studies *was/were* variation.

Finally, we grew a conditional inference tree to uncover how the most important predictors worked together in the data set. This analysis (Figure 6) provided an impressive picture. Individual variation split the community, `Polarity` was only influential among a subset of individuals, differentiation by age was present only for this subset and it was further restricted to affirmative contexts. Interactions of this complexity are difficult to model elegantly in the mixed-effects logistic framework.

Given the overwhelming strength of the `Individual` on variable *was/were*, can we conclude that the story is simply the result of individual variation in York (and perhaps more generally)? There are a number of reasons why this cannot be the primary explanation. Recall that there are pervasive internal constraints involving the contrast between affirmative and negative polarity and an effect of proximity (whether a simple contrast between adjacent/non-adjacent (`Adjacency`) or the influence of a plural element (`Proximate1` or `DP constituency`). The new tools we have used here have demonstrated that each of these predictors are statistically significant over and above the effect of `Individual`, depending on the model. Studies that do not bring `Individual` into the model specification not only run the risk of failing to come to grips with an essential source of variation, they also run the risk of reporting a result as significant which upon closer inspection turns out not to be not significant, i.e. an anti-conservative interpretation of results (see, e.g. Baayen, 2008; Baayen et al., 2008).

In the last step of our analysis we investigated whether and how restricting the analysis to non-categorical individuals might affect our conclusions. It turned out that an analysis of variable individuals only removed `Age` as predictor, while bringing to the fore the effects of `Polarity`, `Adjacency`, `Proximate1` (Figure 7), while supporting the importance of `DP Constituency`. However, we also observed that whether an individual is categorical in her choice of *was* or *were* is predictable from her age, with less variable behavior for younger individuals. For our data set, removal of categorical individuals therefore seems ill-advised, as it introduces a bias against younger individuals in the analysis. For these reasons we do not put much stock in the re-ranking of predictor importance shown for the variable speakers only.

Taken together, these new analyses permit us to offer the following explanation for *was/were* variation in York. The two predictors — `Polarity` and `Adjacency` — actually reflect two different linguistic mechanisms that have separate and independent sociolinguistic repercussions. In affirmative contexts there is language change in progress. It is incremental, roughly linear and steady. We conclude that use of existential *was* is taking its place in the spoken vernacular of English, at least as spoken in northern England at the turn of the 21st century. The fact that the same trajectory of change has been found in real and apparent time in Appalachian English(Montgomery, 1989), Tristan da Cunha English (Schreier, 2002), New Zealand English (Hay and Schreier, 2004) and Australian English (Eisikovits, 1991) supports this interpretation and suggests it extends to other varieties of English. The fact that the DP Constituency comes to the fore when the various predictors involving proximity are tested together exposes an unpredicted result. It suggests that the use of *was* may not be driven by either functional factors or agreement relations but instead involves the syntax of the DP itself.

The effect of polarity is a different process altogether. Based on one of the most productive mechanisms in historical change — morphologization — the use of the *was/were* contrast can encode a polarity contrast rather than agreement. Recall the dramatic picture of affirmative vs. negative in apparent time in Figure 2. Interestingly, closer inspection of the data clarifies that all instances of *was* in negatives were produced by less educated women, e.g. *There wasn't hotels like there is now.* However, none of the models picked this up. This may be due to the very small numbers or the fact that the women simply talk more than the men. The fact that the women also use more *were* than the men provides corroborating evidence. It is interesting that the effect of negation is reported virtually everywhere that *was/were* variation exists; however, the locally enshrined pattern, *There wasn't* or *There weren't*, varies from one speech community to another. The heightened use of a non-standard form among a sociolinguistically salient sector of the population of York suggests an interpretation of this pattern as social, not linguistic.

Thus, we suggest that the products of morphologization can be co-opted to function in the sphere of social meaning to mark particular social groups. This could explain why remorphologization for *was* has been a fundamental part of the explanation for *was/were* variation in North America (e.g., Schilling-Estes and Wolfram, 1994). It may also explain why the the correlation can go either way, more *was* for negatives or more *was* for affirmatives (e.g., Tagliamonte, 2009). We might predict, for example, that if a variety has no effect of negation then it may not have social reallocation of *was/were* variation. Further detailed investigation of patterns of *was/were* variation in contexts of negation will clarify these hypotheses.

In sum, *was/were* variation offers a unique showcase of the primordial drives in linguistic variation and change. The ostensible beginning point for *was/were* variation was a structural agreement relationship governed by syntactic mechanisms of case assignment and hierarchical connection. However, somewhere along the line a stronger force must have challenged the structural agreement bond. The creation of morphological contrasts, which play a central role in grammatical change, was perhaps one of those forces. These appear to be especially amenable to the embedding of social meaning. The tension between agreement rules and linear processing appears to be part of the evolution of this grammatical system and remain immune to social conditioning. In these data, linear processing rather than structural relationship between verb and referent provided a better explanation for the use of *was*. Instead, the internal constituency of the DP, i.e. whether it is construed as singular or plural, may prove more informative. In any case, the results offer several predictions that can now be tested in other speech communities. First, the effect of adjacency as measured by a binary distinction between the verb and plural referent can be expected to negatively correlate with the developmental trajectory of existential *was* such that the effect levels as the frequency of *was* increases. Second, polarity effects, can be expected to correlate with extra-linguistic predictors, although the way a speech community will manifest this effect — if it manifests it at all — will vary. Indeed, these new results for variable (was) suggest more generally that contrasting factors on variable processes may have pointedly distinct interpretations. Thus, the new statistical tools we have employed here may pioneer a whole new type of evidence from which to distinguish the multiplex predictors influencing linguistic variation.

5.1 Conclusion

Let us now return to the issue of methodological practice. Of the models we have considered, the mixed-effects model and the random forest provide the closest fits to the data. In general, the mixed-effects model is an excellent choice for relatively balanced data sets with one or more, potentially crossed, random effect factors (individuals, words, constituents, etc.). For highly unbalanced designs and complex interactions, conditional inference trees and random forests are more flexible, and may yield superior models. However, for large data sets with multiple random-effect factors with many levels, they rapidly become computationally intractable, given current hardware. (Estimating the conditional variable importance for the full data set required approximately 8 hours of processing time on a state-of-the-art CPU.)

Common variationist practice is to code factors (predictors) hypothesized to impact linguistic variables in as elaborated a fashion as possible and then 'hone the analysis' down to the best possible model of the data (e.g., Tagliamonte, 2006). The reason is, of course, the massive covariation across factor groups, empty cells and extreme differences in cell counts typical of analyses of natural speech data. The methodological assistance of a random forest analysis is that it is immune to these problems, allowing the analyst to throw all the factor groups and all the factors into the analysis at the same time and let the analysis evaluate the relative importance of factors. While such a strategy should not be substituted for a linguistically reasoned model, after all the old adage of "garbage in, garbage out" applies nonetheless, it offers the analyst at the very least a preliminary view on the nature of the data set and the impact of the predictors. The conditional inference tree offers yet another perspective since it reveals how interactions and predictors operate in tandem. Indeed, the hierarchical organization of the variable grammar (social and linguistic) is laid out in panoramic relief. Taken together, these new tools can complement and guide the selection of predictors for linear modeling. We conclude that conditional inference trees and random forests, together with mixed-effects models, are practical and effective statistical techniques to add to the sociolinguist's toolkit.

Appendix

Example R code

In this study, we have used R (R Development Core Team, 2009) for the statistical analyses. The simple main effects models presented in this study can be obtained using the variable rule program (Cedergren and Sankoff, 1974), GoldVarb (Rand and David Sankoff, 1990), GoldVarb X (Sankoff et al., 2005), Rvarb (Paolillo, 2002), and Rbrul (Johnson, 2009). Rbrul also allows for straightforward inclusion of interactions and covariates in the model specification. Mixed-effects models require Rbrul or plain R with the lme4 package (Bates and Maechler, 2009). To our knowledge, the conditional inference trees, and random forests based on conditional inference trees, have so far been implemented in R only, in the `party` package. For the following analyses, the `lme4` and `party` packages have to be activated first, as well as the `rws` package in order to have access to the function for calculating the index of concordance C.

```
> library(party)
> library(lme4)
> library(rws)
```

The data are available under the name `york` in the R data frame format on the first author's website, and can be loaded into R as follows:

```
> york =
    read.csv("http://individual.utoronto.ca/tagliamonte/Downloads/york.csv",
    header=TRUE)
```

The simple main effects model of Table 6 and the model including an interaction of Polarity by (nonlinear) Age (Table 5) can be obtained with the following lines of code. The last line carries out an analysis of deviance to ascertain whether the investment in additional parameters by the second model leads to a significantly improved fit to the data.

```
> york.glm1 = glm(Form~Adjacency+Polarity+Age, data=york,
    family="binomial")
> york.glm2 = glm(Form~Adjacency+Polarity*poly(Age,2,raw=TRUE),
    data=york, family="binomial")
> anova(york.glm1, york.glm2, test="Chisq")

Analysis of Deviance Table
Model 1: Form ~ Adjacency + Polarity + Age
Model 2: Form ~ Adjacency + Polarity * poly(Age, 2, raw = TRUE)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       485     631.28
2       482     616.76  3   14.524 0.002273
```

A reasonable mixed-effects model is obtained as follows:

```
> york.lmer = lmer(Form ~ Adjacency + Polarity * poly(Age, 2, raw=FALSE) +
    (1|Individual), data = york, family = "binomial")
> print(york.lmer)
```

A random forest with unbiased conditional inference trees is obtained with

```
> york.cforest = cforest(Form ~ Adjacency + Polarity + Age +
    Sex + Education + Modification + Proximate1.adj + Proximate1 +
    Proximate2 + Individual, data = york)
```

Assessment of the relative importance of the (correlated) predictors requires conditional permutation variable importance, `conditional=TRUE` of the `varimp` function (this requires many hours of processing time with current hardware):

```
> york.cforest.varimp = varimp(york.cforest, conditional=TRUE)
> dotplot(sort(york.cforest.varimp))
```

Assessment of classification accuracy is obtained with `treeresponse`,

```
> york.trp = treeresponse(york.cforest)
> york$PredFOREST = sapply(york.trp, FUN=function(v)return(v[2]))
> york$FormBin = (york$Form=="S")+0
> somers2(york$PredFOREST, york$FormBin)
```

the best single conditional inference tree is produced with:

```
> york.ctree = ctree(Form ~ Adjacency + Polarity + Age +
  Sex  +  Education + Modification + Individual, data=york)
> plot(york.ctree)
```

References

David Adger. Combinatorial variability. *Journal of linguistics*, 42(3):503–530, 2006.

David Adger and Jennifer Smith. Variation and the minimalist program. In L. Cornips and K.P. Corrigan, editors, *Syntax and variation: Reconciling the biological and the social*, pages 149–178. John Benjamins, Amsterdam and Philadelphia, 2005.

David Adger and Jennifer Smith. Language variability and syntactic theory. *UCLA*, 2007.

Lieselotte Anderwald. *Negation in non-standard British English: gaps, regularizations, and asymmetries.* Routledge, 2002.

R. Harald Baayen. *Analyzing Linguistic Data: A practical introduction to statistics using R.* Cambridge University Press, Cambridge, U.K., 2008.

R. Harald Baayen. A real experiment is a factorial experiment? *The Mental Lexicon*, 5:149–157, 2010.

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412, 2008.

Douglas M. Bates. Fitting linear mixed models in R. *R News*, 5:27–30, 2005.

Douglas Bates and Martin Maechler. *lme4: Linear mixed-effects models using S4 classes*, 2009. URL `http://CRAN.R-project.org/package=lme4`. R package version 0.999375-32.

David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics. Identifying Influential Data and sources of Collinearity.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1980.

Theresa Biberauer and Marc Richards. *True optionality: When the grammar doesn't mind.* Department of Linguistics, University of Cambridge, 2008.

Derek Bickerton. Inherent variability and variable rules. *Foundations of Language*, 7:457–92, 1971.

Derek Bickerton. On the nature of a creole continuum. *Language*, 49:640–469, 1973.

Karthryn Bock and Carol A. Miller. Broken agreement. *Cognitive Psychology*, 23:45–93, 1991.

Kathryn J. Bock and Anthony S. Kroch. The isolability of syntactic processing. In Greg N. Carlson and Michael K. Tannenhaus, editors, *The Isolability of Syntactic Processing. Linguistic structure in language processing*, pages 157–196. Kluwer, Dordrecht, 1988.

Kersti Börgars and Carol Chapman. Agreement and pro-drop in some dialects of English. *Linguistics*, 36:71–98, 1998.

Leo Breiman, Jerome H. Friedman, Richard Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, 1984.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

David Britain. Diffusion, levelling, simplification and reallocation in past tense BE in the English Fens. *Journal of sociolinguistics*, 6(1):16–43, 2002.

David Britain and Andrea Sudbury. There's tapestries, there's photos and there's penguins: Variation in the verb BE in existential clauses in conversational New Zealand and Falkland Island English. *Methods*, X, 1999.

David Britain and Andrea Sudbury. There's sheep and there's penguins; convergence, 'drift' and 'slant' in New Zealand and Falkland Ilsand English. In Mari C. Jones and Edith Esch, editors, *Language change; The interplay of internal, external and extra-linguistic factors*, pages 211–240. Mouton de Gruyter, Berlin, 2002.

Henriette J. Cedergren and David Sankoff. Variable rules: Performance as a statistical reflection of competence. *Language*, 50(2):333–355, 1974.

Jack K. Chambers. Social embedding of changes in progress. *Journal of English Linguistics*, 26: 5–36, 1998.

Jack K. Chambers. Dynamic typology and vernacular universals. In B. Kortmann, editor, *Dialectology meets typology: Dialect grammar from a cross-linguistic perspective*, pages 127–145. Mouton de Gruyter, Berlin and New York, 2004.

Jenny Cheshire. *Variation in an English dialect: A sociolinguistic study*. Cambridge University Press, 1982.

Jenny Cheshire, Vivienne Edwards, and Pamela Whittle. Urban British dialect grammar: the question of dialect levelling. *Verbale Kommunikation in der Stadt*, page 67, 1995.

Jenny Cheshire. Syntactic variation and beyond: Gender and social variation in the use of discourse-new markers. *Journal of Sociolinguistics*, 9(4):479–508, 2005.

Donna Christian, Walt Wolfram, and Nanjo Dube. Variation and Change in Geographically Isolated Speech Communities: Appalachian and Ozark English. *Publication of the American Dialect Society*, 72, 1988.

Leonie Cornips and Karen P. Corrigan. *Syntax and variation: reconciling the biological and the social*. John Benjamins, Amsterdam, 2005.

Gaelan Dodds de Wolf. Social and regional differences in grammatical usage in canadian english: Ottawa and vancouver. *American Speech*, 65:3–32, 1990.

William Downes. *Language and society.* Fontana Press, London, 1984.

Edina Eisikovits. Variation in subject-verb agreement in Inner Sydney English. In Jenny Cheshire, editor, *English Around The World: Sociolinguistic Perspectives*, pages 235–256. Cambridge University Press, Cambridge, 1991.

Ralph W. Fasold. *Tense marking in Black English: A linguistic and social analysis.* Center for Applied Linguistics, Washington, D.C., 1972.

Ralph W. Fasold. Tense and the form be in Black English. *Language*, 45(4):763–776, 1969.

Crawford Feagin. *Variation and change in Alabama English: A sociolinguistic study of the white community.* Georgetown University Press, Washington, D.C., 1979.

A. Gilmour, B.J. Gogel, B.R. Cullis, S.J. Welham, and R. Thompson. ASReml user guide, release 1.0. 2002.

Gregory R. Guy. Variation in the group and the individual: the case of final stop deletion. In William Labov, editor, *Locating language in time and space*, pages 1–36. Academic Press, New York, 1980.

Gregory R. Guy. Advanced VARBRUL analysis. In Kathleen Ferrara, Becky Brown, Keith Walters, and John Baugh, editors, *Linguistic Change and Contact*, pages 124–136. Department of Linguistics, University of Texas at Austin, Austin, Texas, 1988.

Frank E. Harrell. *Regression modeling strategies.* Springer, Berlin, 2001.

Jennifer Hay and Danny Schreier. Reversing the trajectory of language change: Subject–verb agreement with be in New Zealand English. *Language Variation and Change*, 16(3):209–235, 2004.

Kirk Hazen. Dialect Affinity and Subject-Verb Concord: The Appalachian Outerbanks. *SECOL Review*, 20:25–53, 1996.

Alison Henry. *Belfast English and Standard English: Dialect variation and parameter setting.* Oxford University Press, New York & Oxford, 1995.

Alison Henry. Parameter setting within a socially realistic linguistics. *Language in Society*, 27: 1–21, 1998.

Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. Survival ensembles. *Biostatistics*, 7:355–373, 2006a.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674, 2006b.

T. Florian Jaeger. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446, 2008.

Daniel Ezra Johnson. Getting off the GoldVarb standard: Introducing Rbrul for mixed effects variabel rule analysis. *Language and Linguistics Compass*, 3:359–383, 2009.

Brian D. Joseph and Richard D. Janda. *The handbook of historical linguistics*. Blackwell, Oxford, 2003.

Brian Joseph and Richard D. Janda. The how and why of diachronic morphologization and demorphologization. In Michael Hammond and Michael Noonan, editors, *Theoretical morphology*, pages 193–210. Academic Press, New York, 1986.

Paul Kay. Variable rules, community grammar, and linguistic change. In D. Sankoff, editor, *Linguistic Variation: Models and Methods*, pages 71–83. Academic Press, New York, 1978.

Paul Kay and Chad McDaniel. On the logic of variable rules. *Language in Society*, 8:151–187, 1979.

William Labov. Contraction, deletion, and inherent variability of the English copula. *Language*, 45 (4):715–762, 1969.

William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, 1972a.

William Labov. The social stratification of (r) in New York City. In *Sociolinguistic Patterns*, pages 43–69. Philadelphia, 1972b.

William Labov, Paul Cohen, Clarence Robins, and John Lewis. A Study of the Non-Standard English Negro and Puerto-Rican Speakers in New York City. Report on Cooperative Research Project 3288. *New York: Columbia University*, 1968.

Beatriz R. Lavandera. Where does the sociolinguistic variable stop? *Language in Society*, 7(2): 171–183, 1978.

Marjory Meechan and Michelle Foley. On Resolving Disagreement: Linguistic Theory and Variation — There's Bridges. *Language Variation and Change*, 6:63–85, 1994.

James Milroy and Lesley MiMilroy. *Real English: the grammar of English dialects in the British Isles*. Addison-Wesley Longman Ltd, 1993.

Gary L. Milsark. Toward an explanation of certain peculiarities of the existential construction in English. *Linguistic Analysis*, 3:1–31, 1977.

MLwiN. MLwiN 2.1. Centre for Multilevel Modeling, University of Bristol, 2007. URL `http://www.cmm.bristol.ac.uk/MLwiN/index.shtml`.

Michael B. Montgomery. Exploring the roots of Appalachian English. *English World-Wide*, 10: 227–278, 1989.

John Nelder. Announcement by the Working Party on Statistical Computing: GLIM (Generalized Linear Interactive Modelling Program). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):259–261, 1975.

John Paolillo. *Analyzing linguistic variation: Statistical models and methods*. CSLI Publications, Stanford, CA, 2002.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

David Rand and David Sankoff. *GoldVarb: A variable rule application for the Macintosh.* Centre de recherches mathématiques, Université de Montréal, Montreal, 1990.

John Rickford. Carrying the new wave into syntax: The case of Black English bin. In R. Fasold and R. Shuy, editors, *Carrying the new wave into syntax: the case of Black English bin. Analyzing Variation in Language*. Washington, D.C., 1975.

Pascale Rousseau and David Sankoff. Advances in variable rule methodology. In D. Sankoff, editor, *Linguistic Variation: Models and Methods*, pages 57–69. Academic Press, New York, 1978.

David Sankoff. Sociolinguistics and syntactic variation. *Linguistics: the Cambridge Survey*, 4: 140–161, 1988.

David Sankoff. *Linguistic variation: Models and methods.* Academic Press, New York, 1978a.

David Sankoff. Probability and linguistic variation. *Synthèse*, 37:217–238, 1978b.

David Sankoff. Probability and linguistic variation. *Synthèse*, 37:217–238, 1978c.

David Sankoff. Sociolinguistic method and linguistic theory. In L. Jonathan Cohen, Jerzy Los, Helmut Pfeiffer, and Klaus Peter Podewski, editors, *Logic, methodology, philosophy of Science VI*, pages 677–689. North Holland, Amsterdam, 1982.

David Sankoff. Statistics in linguistics. In *Encyclopaedia of the statistical sciences.* Wiley, New York, 1985.

David Sankoff and Suzanne Laberge. The linguistic market and the statistical explanation of variability. In D. Sankoff, editor, *The Linguistic Market and the Statistical Explanation of Variability. Linguistic Variation: Models and Methods*, pages 239–250. New York, 1978.

David Sankoff and William Labov. On the uses of variable rules. *Language in Society*, 8:189–222, 1979.

David Sankoff and Pascale Rousseau. Categorical contexts and variable rules. In S. Jacobson, editor, *Papers from the Scandinavian Symposium on Syntactic Variation, Stockholm, May 18-19, 1979.*, pages 7–22. Almqvist and Wiksell, Stockholm, 1979.

David Sankoff and Gillian Sankoff. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In R. Darnell, editor, *Canadian Languages in their Social Context*, pages 7–63. Linguistic Research Inc., Edmonton, 1973.

David Sankoff, Sali A. Tagliamonte, and Eric Smith. *Goldvarb X.* Department of Linguistics, University of Toronto, Toronto, Canada, 2005. URL `http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm`.

David Sankoff, Sali A. Tagliamonte, and Eric Smith. *Goldvarb Lion*, 2012. URL `http://individual.utoronto.ca/tagliamonte/goldvarb.htm`. Department of Linguistics, University of Toronto, Toronto, Canada.

Gillian Sankoff. Cross-sectional and longitudinal studies in sociolinguistics. In Ulrich Ammoon, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill Trudgill, editors, *Inernational handbook of the science of language and society*, pages 1003–1013. Mouton de Gruyter, Berlin, 2005.

Natalie Schilling-Estes and Walt Wolfram. Convergent explanation and alternative regularization patterns: Were/weren't leveling in a vernacular English variety. *Language Variation and Change*, 6:273–302, 1994.

Danny Schreier. Past be in Tristan da Cunha: The rise and fall of categoricality in language change. *American Speech*, 77(1):70, 2002.

Carolin Strobl, James Malley, and Gerhard Tutz. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, 14(4):323–348, 2009.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 2007. URL `http://www.biomedcentral.com/1471-2105/8/25`.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 2008. URL `http://www.biomedcentral.com/1471-2105/9/307`.

Sali A. Tagliamonte. Was/were variation across the generations: View from the city of York. *Language Variation and Change*, 10:153–191, 1998.

Sali A. Tagliamonte. Come/came variation in English dialects. *American Speech*, 76:42–61, 2001.

Sali A. Tagliamonte. Comparative sociolinguistics. In J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, editors, *Handbook of language variation and change*, pages 729–763. Blackwell Publishers, Malden and Oxford, 2002a.

Sali A. Tagliamonte. Variation and change in the british relative marker system. In Patricia Poussa, editor, *Relativisation on the North Sea Littoral*, pages 147–165. Lincom Europa, Munich, 2002b.

Sali A. Tagliamonte. "every place has a different toll": Determinants of grammatical variation in cross-variety perspective. In Gunter Rhodenberg and Britta Mondorf, editors, *Determinants of grammatical variation in English*, pages 531–554. Mouton de Gruyter, Berlin and New York, 2003.

Sali A. Tagliamonte. *Analysing sociolinguistic variation*. Cambridge University Press, Cambridge, 2006.

Sali A. Tagliamonte. There was universals; then there weren't: A comparative sociolinguistic perspective on 'default singulars'. In Markku Fillpula, Juhani Klemola, and Heli Paulasto, editors, *Vernacular Universals versus Contact Induced Change*, pages 103–129. Routledge, Oxford, 2009.

Sali A. Tagliamonte and Rebecca V. Roeder. Variation in the English definite article: Socio-historical linguistic in t'speech community. *Journal of Sociolinguistics*, 13:435–471, 2009.

Sali A. Tagliamonte and Jennifer Smith. Analogical levelling in samaná english: the case of was and were. *Journal of English Linguistics*, 27:8–26, 1998.

Sali A. Tagliamonte and Jennifer Smith. Old was; new ecology: Viewing English through the sociolinguistic filter. In S. Poplack, editor, *The English history of African American English*, pages 141–171. Blackwell Publishers, Oxford and Malden, 2000.

Sali A. Tagliamonte and Jennifer Smith. Layering, change and a twist of fate: Deontic modality in dialects of English. *Diachronica*, 23:341–380, 2006.

Peter J. Trudgill. *The dialects of England*. Blackwell Publishers, Oxford, 1990.

Hans van de Velde and Roeland van Hout. Dangerous aggregations. a case study of dutch (n) deletion. In Carole Paradis, editor, *Papers in Sociolinguistics*, pages 137–147. Nuits Blanches, Quebec, 1998.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S-Plus*. Springer, New York, 4 edition, 2002.

James A. Walker. "There's bears back there": Plural existentials and vernacular universals in (Quebec) English. *English World-Wide*, 28(2):147–166, 2007.

Brady T. West, Katheleen B. Welch, and Andrzeij T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC Press, 2007.

Walt Wolfram and Donna Christian. *Appalachian Speech*. Center for Applied Linguistics, Arlington, Virginia, 1976.

Walt Wolfram. A Sociolinguistic Description of Detroit Negro Speech. *Urban Language Series*, 5, 1969.

Walt Wolfram. Identifying and interpreting variables. In D. Preston, editor, *American dialect research*, pages 193–221. John Benjamins, Amsterdam and Philadelphia, 1993.