This is an Accepted Manuscript of a book chapter published by Routledge in Social and Regional Variation in World Englishes: Local and Global Perspectives on 30 September 2022, available online: https://doi.org/10.4324/9781003227342

4. *Was/were* Variation with Subject Pronouns *We, You,* and *They* in Recent British English – Towards Standard Uses?

Paula Rautionaho
Orcid 0000-0002-5239-8407
Mark Kaunisto
Orcid 0000-0002-0084-1067

Abstract

This chapter explores trends in the *was*/*were* variation in recent British English, focussing on the uses of *was*/*were* with the pronoun subjects *we*, *you*, and *they* in the spoken demographic part of the BNC and the Spoken BNC2014 corpus. Extracting all instances of *was* and *were* with the examined pronoun subjects in both corpora, we annotated the data for intra-linguistic (e.g. negation, pronoun) and sociolinguistic (age, gender, region, and social class) variables. While a striking decline in the normalized frequencies of *was* with the pronouns is undisputable, we dig deeper into intra- and extra-linguistic parameters to reveal the changing patterns with generalized linear mixed model tree analysis (GLMM tree; Fokkema et al. 2018). The results indicate that the sociolinguistic parameters override intra-linguistic ones; the major divide is found between speakers of different age groups, working class speakers as opposed to other social classes, and the north and the south, while pronouns and inversion are the only intra-linguistic parameters selected in the final model.

1. Introduction

From a morphosyntactic point of view, the verb *be* has the most complex system of indicative verb forms in the English language, with separate forms being used according to the person and number of the subject. There are three different present tense forms (*am*, *are*, *is*), and two past tense forms (*was*, *were*). What further contributes to the complexity is that some forms are used for only one particular combination of person and number of the subject (*I am, he*/*she*/*it is*), whereas other forms are used with different kinds of subjects in this regard *(we*/*you*/*they are*; *I*/*he*/*she*/*it was*; *we*/*you*/*they were*). Even as regards the use of the two past tense forms, there has been variation in the ways that *was* and *were* have been used, and in recent years, an increasing number of studies (see e.g., Tagliamonte 1998; Anderwald 2001, 2002; Schreier 2002; Cheshire and Fox 2009) have focussed on the *was/were* variation and the factors that influence non-standard uses of *was* (as in (1)) and *were* (as in (2)). Many studies approach the variation from a sociolinguistic perspective, and factors relating to different speaker groups (age, region, ethnicity, gender, etc.) have been identified and

weighed in terms of relative degrees of adherence or non-adherence to standard uses of the verb forms. In addition to such language-external factors, the variation has also been observed to manifest itself differently depending on the morpho-syntactic characteristics of the immediate linguistic context, such as the types of subjects or clause negation having an effect on the likelihood to use non-standard *was* or *were*.

(1) That's what **we was** up against. (BNC1994DS, KBB)
(2) **She were** coughing like mad! (BNC1994DS, KB1)

As will be observed in Section 2, the variation in the use of *was* and *were* offers a rich field for linguistic investigation for a number of reasons. To begin with, variation has been attested at different periods in the history of English – according to Tagliamonte (1998), evidence is found of *was*/*were*-variation already in Old English – and different forms of non-standard uses have emerged at various times in different speaker communities. As far as the spoken vernacular is concerned, the use of *was* and *were* is in a constant state of flux to varying degrees in different parts of the world. In addition to temporal and regional dimensions, the nature of available data and methodological considerations allow us to approach the question from different angles. Impetus to explore the matter further comes from the available data in the BNCSpoken2014 corpus, which together with the original *British National Corpus* offers possibilities of studying two large collections of spoken data, compiled one generation apart, and both including data representing speech of British people of different age groups, regions, social classes, factors also examined in earlier studies. It is of interest to study the factors most commonly associated with non-standard *was* in the two sets of data, and to what extent the situation has changed.

Through the years, the methods of compiling data and the analytical tools applied in the study of *was*/*were* variation have also changed. Some studies have involved traditional fieldwork with interviews, which have then been transcribed by the authors themselves, while other studies have been based on large electronic and grammatically tagged databases allowing complex search queries. The increased corpus sizes have made it possible to conduct multivariate analyses in order to shed more light on which language-internal and language-external factors appear to have been most significant in affecting non-standard uses of *was* and *were*; for example, Tagliamonte and Baayen (2012) examined *was*/*were*-variation in plural existential constructions in the materials from the city of York – studied previously in Tagliamonte (1998) – by making use of tools including mixed-effects models, random forests, and conditional inference trees. Using data from the

demographically sampled part of the spoken section of the *British National Corpus* (henceforth referred to as BNC1994DS) and BNCSpoken2014, the present chapter aims to examine the applicability of modelling methods such as generalized linear mixed model (GLMM) tree analysis (Fokkema et al. 2018) on the data regarding non-standard uses of *was* with the pronouns *we*, *you*, and *they*, as in *we*/*you*/*they was late for the party*. The overall questions that the study aims to shed light on are the following:

i. What diachronic trends can be observed with regard to the *was* vs. *were* alternation with the examined pronoun subjects?
ii. To what extent do age, gender, social class, and dialect area govern the speaker's choice of *was* over *were*?
iii. To what extent do the pronoun, inversion, or polarity govern the choice?

The second section of this chapter looks into previous studies on *was*/*were* variation attested in various vernaculars of English around the world, concentrating on the different kinds of factors that have been observed to play some role in the types and degrees of variation in the use of the two past tense forms. This will be followed by sections describing the characteristics of the data and the steps that were taken in its analysis, as well as presentation and discussion of the main findings. As will be seen, the results reveal clear trendlines of overall change, with a dramatic decrease in the non-standard use of *was* between the BNC1994DS and the BNCSpoken2014 data. Some evidence can be perceived of differences relating to the speakers' age, regional dialect, sex, and social class. However, the reality of diminishing numbers of occurrences of non-standard *was* across the board inevitably presents us with challenges in determining the most important factors influencing its use: in other words, when fewer people altogether are using the non-standard form, it becomes harder to assess the significance of multiple factors as influencing the change over time.

2. Previous research

Was/*were* variation has been studied from a number of angles and with different focal points. As far as the geographical dimension is concerned, studies have ranged from ones focussing on smaller communities, towns, cities, or islands (e.g., Cheshire 1982, Tagliamonte 1998, Smith and Tagliamonte 1998, Schreier 2002, Cheshire and Fox 2009, and Durham 2013), regions (e.g., Britain 2002) to investigations on a nationwide scope of analysis (e.g., Anderwald 2001, 2002; Hay and Schreier 2004). Mostly the studies have concentrated on contemporary vernacular speech, but historical periods and written language have also been examined, including, for example, Nevalainen (2006) on *was*/*were* variation in Late Middle and Early Modern English letters.

Even though the use of the past tense forms of *be* deals with only two items, *was* and *were*, the non-standard uses of past tense *be* have manifested themselves at various points in time in different ways around the world. Basically, non-standard uses either involve the use of *was* in instances where *were* would normally be used (i.e., with plural subjects, as in (1) above, as well as with singular *you*), or the use of *were* instead of the standard *was* (i.e., with singular subjects, as in (2) above).[1] The first type of non-standard use is often referred to as *was*-levelling, and the latter as *were*-levelling (or *was*- and *were*-generalization, respectively). Furthermore, as has been shown in previous studies, one factor adding to the complexity of the system and to the numbers of types of non-standard uses is negation, as in some dialect areas or speech communities *wasn't* and *weren't* are not necessarily used in the same ways as *was* and *were*. In fact, there have been reports of non-standard uses involving *was*-levelling in positive contexts whereas in negated contexts the use of *weren't* has gained ground, particularly in tags (e.g., Tagliamonte 1998, 179; Anderwald 2002, 178). In other words, in some dialects *was* has become the preferred form with both singular and plural subjects, but in negative contexts and again regardless of the number of the subject, *weren't* is the preferred verb form (i.e., *weren't I*/*it*/*we*/*you*/*they*).

Although different types of levelling have been observed – and the degrees of non-standard uses of *was* and *were* are typically reported in terms of percentages out of all uses of the two forms – *was*-levelling has been regarded as being the more frequently occurring type (see e.g., Wolfram and Schilling-Estes 2003, 132). Such observations have led Chambers (1995; 2004, 136) to more emphatically consider invariant *was* as the basic pattern of non-standard past tense forms of *be*, constituting even a "vernacular root" used as a default option (see also Cheshire and Fox 2009, 3). This claim has been supported by findings that instances of *was*-levelling appear to go further back in history, whereas trends of *were*-levelling are of more recent origin (Britain 2002; Nevalainen 2006), as well as observations of post-colonial Englishes showing stronger tendencies of *was*-levelling compared to non-standard uses of *were* (Schreier 2002, 74). Nevertheless, next to the use of non-standard *was* in both positive and negative contexts as the major vernacular pattern, there is also the previously mentioned *was*/*weren't* pattern with polarity as an important factor affecting the choice of verbal form. This pattern has also been found in various dialects around the world (e.g.,

[1] Another point worth noting is the increasing use of *they* as a gender-neutral pronoun with singular reference; as regards the present study, however, the tokens of *they* in the data were not examined in detail in terms of whether they had singular or plural reference. Our general impression was that instances of gender-neutral pronoun uses of *they* in the data examined were not very frequent.

Tagliamonte (1998) in York, Britain (2002) in the Fens, and Anderwald (2001) in several dialect areas across the United Kingdom).

Many studies have also observed the variability of the strength of the *was*/*were*-levelling over time, with some generations showing greater tendencies over the use of different non-standard uses. In her study on the *Corpus of Early English Correspondence*, Nevalainen (2006, 359) saw that the use of non-standard *was* with plural subjects reached its peak in different regions at different times between 1440 and 1681, and overall, this type of use of *was* declined towards the seventeenth century. As regards more recent times in the city of York, Tagliamonte (1998) noted that the use of *was* with plural subjects was likewise showing signs of weakening, while the use of *weren't* with all kinds of subjects was increasing. Similar trends in London were perceived by Cheshire and Fox, who make a point about the "ongoing innovation and change" (2019, 34) that characterizes *was/were* variation.

The likelihood of non-standard uses is to a notable degree also linked to the type of subject itself. Based on a number of studies, there is wide agreement that non-standard *was* appears to occur most likely in existential plural constructions with *there* (as in *there was two sets of keys*), but dialects differ as regards types of plural subjects which are most resistant to the use of the singular form of the verb. Chambers (2004) posits that the pronoun *they* is the least likely to allow or attract *was*, and he suggests the sequence of "*they* – non-pronominal plural nouns – *we* – *you* – existentials" as showing increasing likelihood of permitting non-standard *was*. Tagliamonte's (1998) results from York, on the other hand, show a slightly different sequence, with *we* and *they* together manifesting the lowest degrees of *was*-levelling, followed by plural nouns or noun phrases, then *you*, and finally existential plural constructions.[2] The role of the subject has also been seen to be relevant in connection with variation in tag questions, with Tagliamonte (1998), for instance, reporting on the increase of non-standard *weren't* particularly with the pronoun *it* as the subject.

Studies on the combinations of different language-internal and language-external factors have enabled scholars to examine the ways in which changes take place, what kind of social dynamics are at play, and how the changes actually spread (i.e., which social groups seem to be at the

[2] Related to these language-internal constraints and differences between dialects in Britain is the so-called Northern Subject Rule, which in the case of past tense forms of *be* would indeed show higher degrees of non-standard *was* with plural noun phrases than with plural pronouns, as noted by e.g. Britain (2002, 20) and Cheshire and Fox (2009, 6). For a discussion on the origins of the rule, see e.g. Klemola (2000).

forefront of different types of changes). In broader terms, the studies have contributed to the knowledge of the phenomena related to dialect contacts or even language contacts. An interesting case in point in this regard is the study by Cheshire and Fox (2009), who examined *was*/*were* variation in conversations among adolescent and elderly inner East London as well as outer London speakers. They found varying degrees of the use of non-standard *was*/*weren't* patterns, that is, *was*-levelling in affirmative, and *weren't*-levelling in negative polarity contexts. However, *was*-levelling was less frequent among adolescent inner London speakers. When it came to the types of subjects with which different speaker groups used these non-standard forms, some differences were found. For example, non-standard *was* typically manifested itself with plural NPs, but not with *they* with inner London speakers, whereas in outer London areas examined, the situation was reversed. In addition, non-standard *you was* was frequently attested in the speech patterns of inner London elderly speakers as well as outer London adolescents, whereas outer London elderly people did not use this type of non-standard *was* at all (Cheshire and Fox 2009, 10). They also found that in inner London, the ethnicity of the younger speakers plays a strong role in the frequencies of non-standard *was* as well as *wasn't*, with for instance Afro- Caribbeans using non-standard *wasn't* frequently, as opposed to Bangladeshi speakers, who favoured standard patterns (ibid., 24). Overall, their findings led them to conclude that the variation in London suggests a complex phenomenon with a number of possible ongoing trends involving dialect contacts, social networks and social integration, the comprehensive description of which requires further study.

The closest of the previous studies to ours, especially in terms of the data examined, is that of Anderwald (2001), which focussed on *was*/*were* variation with existential *there* constructions and personal pronoun subjects in spoken data of the *British National Corpus*. Anderwald observed that in affirmative clauses the use of non-standard *was* with these subject types was used by speakers in all regions across the UK, with an average of 12% of all the relevant occurrences having *was* instead of *were*. The generalization of *was* in these contexts manifested itself in varying degrees from one region to another, with Humberside being the region with the lowest rate of non-standard *was* (3%), and East Anglia topping the list in this regard (40%). The situation was notably different in negated clauses, as the use of non-standard *wasn't* was relatively rare (it was used in 5% of all the relevant cases with plural subjects), whereas non-standard *weren't* was used more frequently with singular first and third person subjects (28% of the cases). Again, Humberside and East Anglia were at the opposite ends as regions showing lowest and highest degrees of non-standard uses of *wasn't* and *weren't.* Considering the present study, it is important to note that as we are examining

was/were variation with *we*, *you*, and *they* as subjects, our study comments on the factors of the use of non-standard *was* instead of *were*.

Against this background, we expect to find decreasing frequency of *was*-levelling in our data representing recent BrE. We also expect to see variation pertaining to polarity and subject type, so that *was* is not expected to be the default choice in negative contexts, nor with the subject *they*. Finally, based on earlier research, we anticipate high levels of variation with regard to sociolinguistic parameters such as region and age. The following section presents our data and methods.

3. Data and methods

The study for the present chapter relies on data drawn from the *British National Corpus*, both the original 1994 version and the new 2014 version (Love et al. 2017). We focus on the comparable parts of the two corpora: from BNC1994, we select the demographically sampled part (BNC1994DS), which has approximately 4.2 million words from over 1,000 speakers who were "selected by age group, sex, social class and geographic region" and recorded in the late 1980s and early 1990s (Aston and Burnard 1998, 31). From the BNCSpoken2014, which has material recorded between the years 2012 and 2016 (Love et al. 2017), we select the sample released in 2016 (BNCSpoken2014), which contains approximately 4.8 million words (ibid., 9–10). The two corpora were searched[3] for all instances of *was* occurring with the personal pronouns *we*, *they*, and *you*, either following or preceding the pronouns, and in their negated forms. In order to investigate the alternation between *was* and *were* in contexts where *were* would be the standard choice, we also extracted all instances of *were* occurring with the same pronouns (see Table 4.1 for the exact search phrases).

Table 4.1. Search phrases used.

	was	*were*
Regular (returns also Negated)	"(we|they|you) was"	"(we|they|you) were"
Inverted	"was (we|they|you)"	"were (we|they|you)"

[3] Data cited herein have been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved. We used BNCWeb (bncweb.lancs.ac.uk) for extracting the data from BNC1994DS, and CQPweb (Hardie 2012) for extracting the data from BNCSpoken2014.

Inverted and Negated	"was n't (we|they|you)"	"were n't (we|they|you)"

For the present analysis, all instances of *was* and *were* were manually checked for relevance and relevant tokens retained in the database; we categorically excluded all tokens for which one or more of the four sociolinguistic metadata were missing. Additionally, tokens excluded from the analysis include unclear or incomplete tokens (as in (3) and (4)), or instances where the pronoun and *was/were* actually belong to different phrases (as in (5)).

(3) […] all I know is that **we was** [unclear] first […] (BNC1994DS, KDA)
(4) **they was** they had been sold at full price so […] (BNCSpoken2014, SDHB)
(5) Yeah, but he was a bastard to **you was**n't he? (BNC1994DS, KCU)

Relevant tokens of *was* occurring where *were* would be expected in standard BrE are illustrated in sentences (6) to (8). In (6), *was* occurs as a copula in regular word order, and in (7) in inverted order. Sentence (8) showcases a progressive occurrence, as well as *was* occurring negated and as a tag question.

(6) […] **they was** all muddy […] (BNCSpoken2014, SXDQ)
(7) Where **was you** then? (BNC1994DS, KBF)
(8) **we was** talking about that earlier **wasn't we**? (BNCSpoken2014, SPHU)

Based on earlier research (see Section 2), the instances were annotated for three intra-linguistic variables: INVERSION (no inversion vs. question vs. tag), POLARITY (positive vs. negative), and PRONOUN (*we* vs. *they* vs. *you*). With regard to the pronoun *you*, we decided to keep also the singular form in the dataset as it too requires *were* in standard varieties of English; the instances of *you* were not, however, annotated further in an effort to manage the statistical analysis with the low raw frequencies in BNCSpoken2014. To exemplify the annotation scheme, the tagging sequence for sentence (6) above is "no inversion + positive + *they*", and for (7) it is "question + positive + *you*".

Both BNC corpora include a wealth of metadata on the speakers' occupation and geographical origin, among other parameters. For the present study, we focus on the speakers' age, sex, social class, and dialect area. Unfortunately, not all sociolinguistic metadata provided by the two corpora are readily or fully comparable, which is why we modified the information slightly to obtain a comparable dataset. The modifications involved collapsing some factor levels to arrive at a more

balanced dataset; for instance, the dialect area in BNC1994DS was collapsed to the following three: 'Midlands' vs. 'North' vs. 'South' (see Appendix 1 for details). To tackle the problematic age categories in the two corpora, we retrieved the exact age of the speakers in BNC1994DS and then implemented a simplified categorization following Säily et al. (2018), thus including the following factor levels: '0–29' vs. '30–49' vs. '50–99'. With regard to SOCIAL class, we again followed Säily et al. (2018) and collapsed levels A, B and C1 into 'middle', and C2, D and E into 'working'. Variable SEX has the levels 'female' vs. 'male'. After the data extraction, manual checking and annotation of variables, the dataset consists of 10,669 instances of *was* and *were* in BNC1994DS and BNCSpoken2014 (see Section 4.1 for more detailed analysis).

To gauge the effect of the three intra-linguistic and the four sociolinguistic variables on the choice of *was* over *were* in recent BrE, we ran a generalized linear mixed methods tree analysis (GLMM tree; Fokkema et al. 2018; Fokkema, Edbrooke-Childs and Wolpert 2020) in RStudio (RStudio Team 2021). GLMM trees are, in essence, a combination of the recursive partitioning of a tree-based method (such as conditional inference trees and random forests) and a GLMM, accounting for random-effects parameters. According to Fokkema et al. (2018), each terminal node in a GLMM tree "is associated with different fixed-effects regression coefficients while adjusting for global random effects (such as a random intercept)". Building on GLM trees which consist of subgroup-specific GLMs embedded in the nodes of the tree (Fokkema et al. 2020), GLMM trees are able to account for a multilevel structure of a dataset; in our case, the clustering in the data arises from the individual speakers preferring either *was* or *were*. The binary dependent variable in our analysis is WASWERE (i.e. *was* vs. *were*), the independent variables are CORPUS, INVERSION, POLARITY, PRONOUN, and AGE, SEX, REGION and CLASS, with SPEAKER as the global random variable. While GLMM trees are based on GLMM modelling, there is no need to build the model in a step-wise manner; rather, the tree analysis automatically disregards variables that bear no statistical significance in the model, and also accounts for interactions of variables. The global random effect accounts for any speaker-specific effect: for instance, an individual speaker may be inclined to use *was* clearly more frequently than another one. For the statistical analysis, we include all instances of *was*, and a random sample of *were*, determined with the help of a confidence level of 95% and a margin of error being 3%[4]: the database thus consists of 2,126 tokens, of which 326 tokens represent *was* and 1,800 *were*.

[4] Sample size was calculated with the help of qualtricks, available at https://www.qualtrics.com/blog/calculating-sample-size/ (accessed 29 Dec 2021).

4. Results

4.1 Descriptive statistics

Before presenting the results of the statistical modelling, we first briefly summarize our findings on the individual predictor variables, starting with the intra-linguistic ones and then moving on to the sociolinguistic parameters. Starting with the overall frequency of *was* occurring with the pronouns *we*, *they*, and *you*, we find a drastic decline between BNC1994DS and BNCSpoken2014: in the early 1990s, *was* occurred with the three pronouns in a standard *were* context with a proportion of 7.2% (RF=219), whereas in the 2010s, the corresponding proportion is 1.1% (RF=79; LogLikelihood[5] 234.51). This dramatic drop in the proportion of *was*-levelling overall is striking, and it suggests that the evidence found by Tagliamonte (1998, 184) on the weakening of non-standard *was* in York was probably already signalling broader trends of change in *was*-levelling in various parts of the UK. As Figure 4.1 portrays, the proportion of *was* in standard *were* contexts has decreased rather evenly in regard to the three pronouns investigated: *we* remains the most prominent pronoun to co-occur with *was*, while the proportion of *you* co-occurring with *was* has decreased somewhat more rapidly. *They was* is the least frequent combination in both datasets. These findings do not fully conform to Chambers (2004) and Tagliamonte (1998) in that *we* occurs with *was* more frequently than expected based on these earlier studies.

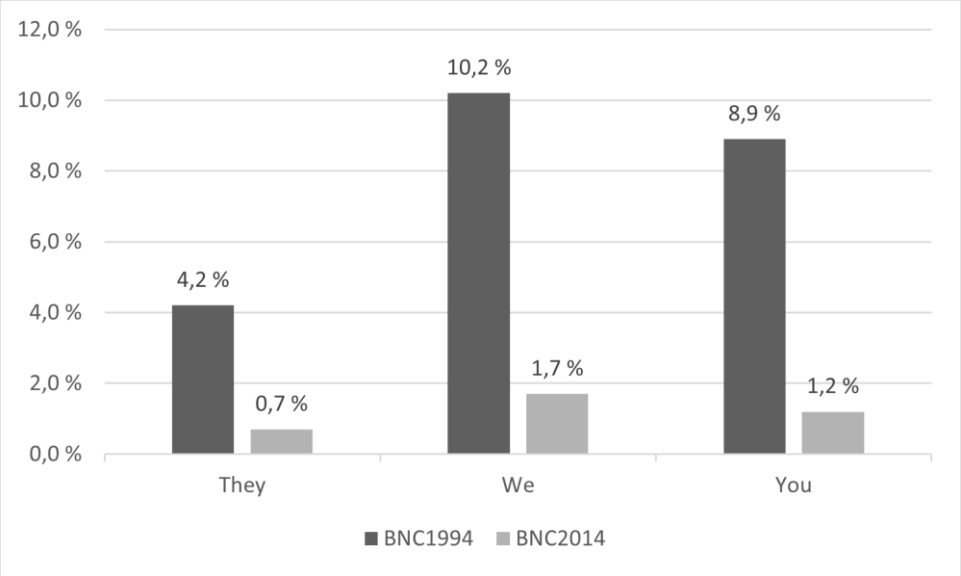

Figure 4.1. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to subject pronoun.

[5] Log likelihood values were calculated with the Log-likelihood and effect size calculator available at http://ucrel.lancs.ac.uk/llwizard.html.

With regard to the intra-linguistic variables included in the analysis, overall, we find statistically significant decrease for almost all factor levels, but there is no single variable that would explain the decrease in the frequency of *was*-levelling. The only factor level showing decrease that is not statistically significant is tags, where the proportion of *was* occurring in standard *were* contexts has not diminished as much as for many other factor levels between the two datasets; Figure 4.2 also shows that in the 1990s, *was*-levelling is most prominent in questions and in contexts with no inversion, but that in the newer dataset there is very little variation detected in this respect. With regard to polarity, Figure 4.3 clearly portrays that in the 1990s, *was* prefers positive contexts, as reported earlier by Tagliamonte (1998) and Anderwald (2002), but that this difference is again levelled out in the newer dataset.

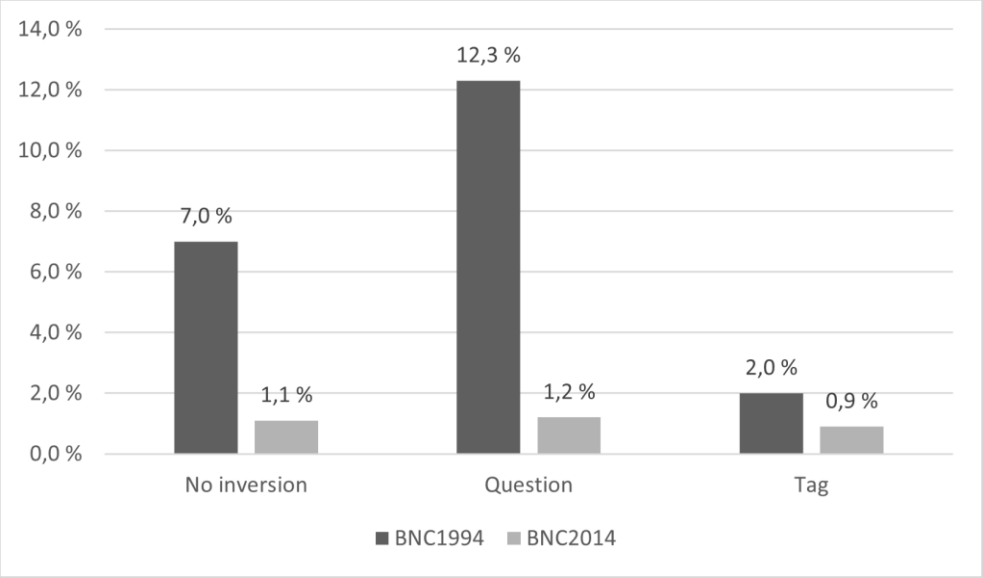

Figure 4.2. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to inversion type.

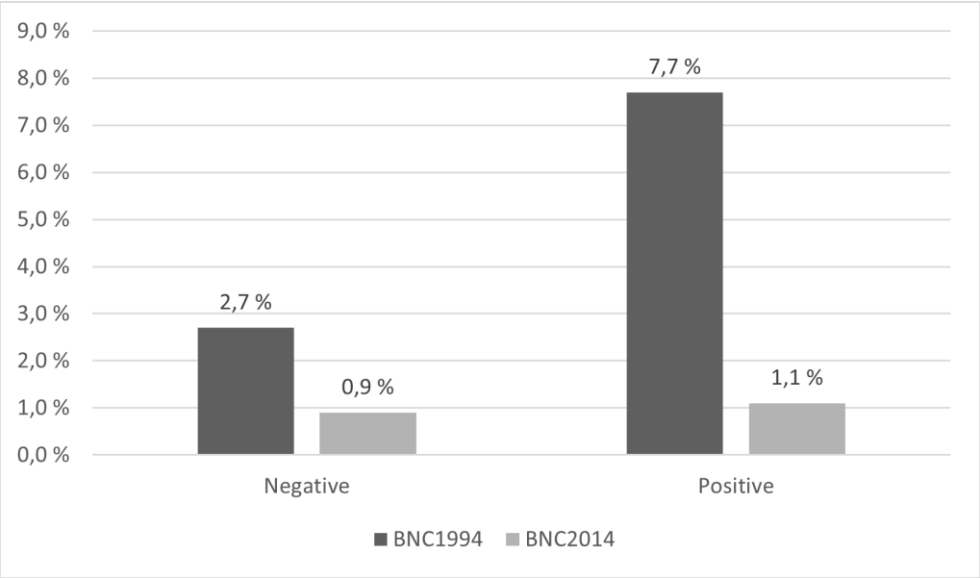

Figure 4.3. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to polarity.

Turning next to the sociolinguistic parameters, some diverging diachronic trends appear, especially with regard to the age and dialect area of those speakers who use *was* in *were*-contexts. Figure 4.4 shows that the proportions of *was*-levelling across different age groups are rather different in the two datasets; in BNC1994DS we find *was*-levelling in all age groups, most frequently by those between the ages of 30 to 49, with the lowest proportions in the youngest and the oldest age groups. In BNCSpoken2014, on the other hand, speakers under the age of 50 rarely use *was* in standard *were* contexts at all, and compared to the earlier dataset, *was*-levelling is clearly less frequent also across the older speakers in the dataset. Overall, the results may represent a case of age gradience in that the high proportion of *was*-levelling in the 30–49 group in BNC1994DS is reflected in the speakers above the age of 50 in BNCSpoken2014. Figure 4.5 shows a subtle shift in the gender patterns; male speakers use non-standard *was* proportionately more frequently in the newer dataset.

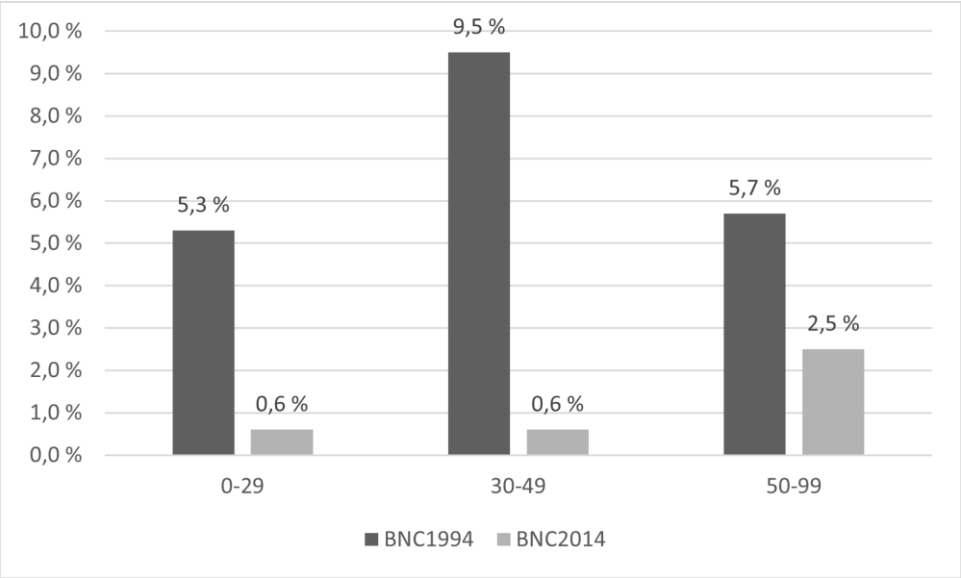

Figure 4.4. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to age group.

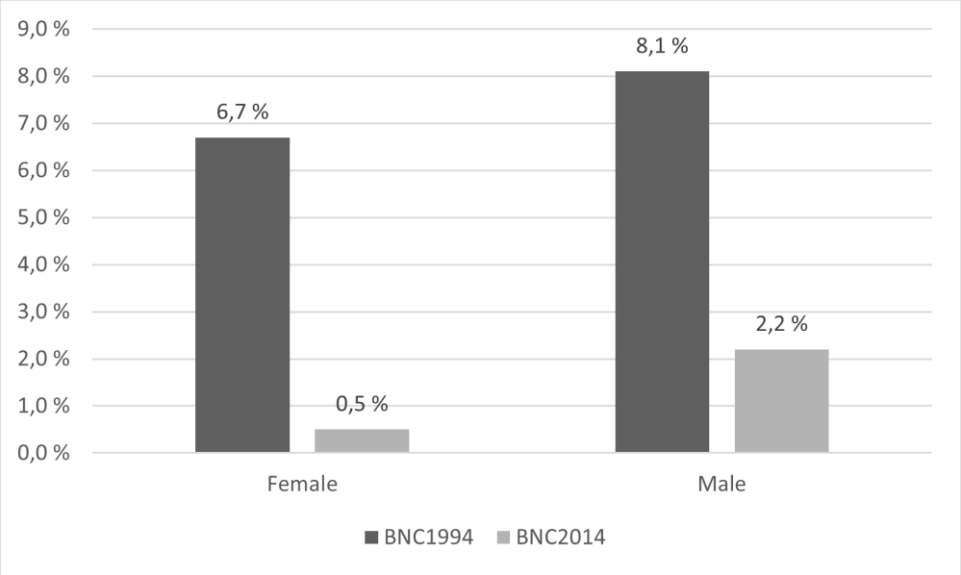

Figure 4.5. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to sex.

Turning next to variation according to social class (see Figure 4.6), we find a general trend of *was*-levelling being more common in the working class as opposed to the middle class in both datasets. In earlier literature, rather few observations have been made on the role of social class in *was*/*were* variation, with age and dialect usually being stronger factors connected with non-standard uses, but

as Anderwald (2002) notes, challenges in the representativeness of data tends to pose difficulties when trying to examine the influence of a number of factors in combination.[6]

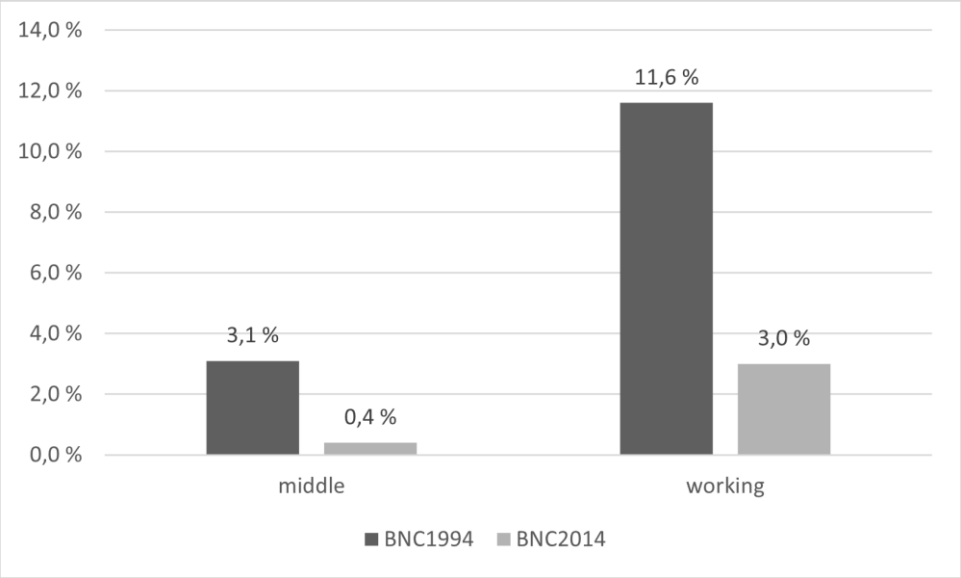

Figure 4.6. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to social class.

Finally, Figure 4.7 indicates geographical shift in the use of *was*: in the earlier dataset, *was*-levelling is most frequently found in the South (11%) and the Midlands (5%), whereas in the newer dataset the frequencies have diminished to less than 1% in both regions. Interestingly, the North has transitioned from a region with the lowest proportion of *was*-levelling in BNC1994DS to the one with the highest in BNCSpoken2014; in other words, the North has retained *was*-levelling and, in the same process, has become the most prominent area of non-standard *was* in the 2010s.

[6] Some observations have been made on different types of *was*/*were* variation connected with the social class of the speakers; for example, Schreier (2002, 75) mentions how Feagin (1979) found *was*-levelling to be least frequent with the pronoun *they* among working class informants.

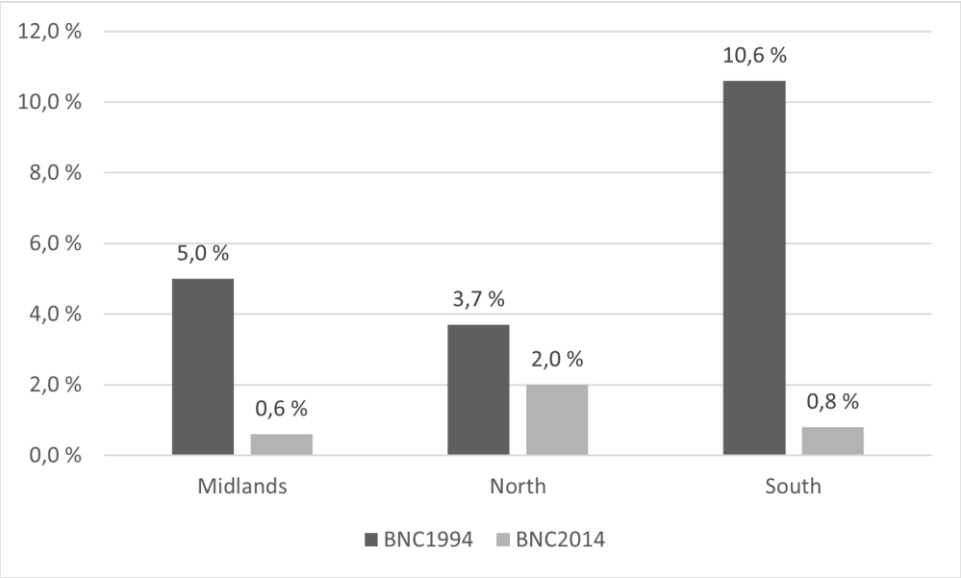

Figure 4.7. Proportion of *was*-levelling in standard *were* contexts in BNC1994DS and BNCSpoken2014, according to geographical region.

To summarize the descriptive statistics presented above, we found an all-encompassing trend of decreasing use of *was*-levelling in the two datasets; most factor levels of the different variables show statistically significant decrease, the only exceptions being tag questions and the North where the decrease is not statistically significant. There appears to be somewhat more variation within the BNC1994DS dataset compared to the BNCSpoken2014 dataset, in which *was* occurs within a tighter envelope of sociolinguistic variation consisting of mostly older speakers and the North. Any differences in the proportions of the intra-linguistic variables (PRONOUN, POLARITY, INVERSION) in BNC1994DS have largely levelled out in BNCSpoken2014. We now turn to the statistical modelling to see how the different variables interact with one another with regard to the choice of *was* over *were*.

4.2 Variable selection: was *over* were

The dataset submitted to the GLMM tree analysis consists of 1,978 tokens, of which 298 represent *was* and 1,680 represent *were*. The GLMM tree modelling was controlled for the optimizer and for the maximum number of iterations.[7] Figure 4.8 presents the resulting tree diagram (C index is 0.94 which is above the level of good performance; see Tagliamonte and Baayen 2012, 204); the light grey bars in the terminal nodes (at the bottom of the figure) illustrate the predicted probability of *was*, as opposed to *were* in dark grey, in the different combinations formed by the predictor

[7] gt <- glmertree(WasWere ~ 1|Speaker|Corpus+Polarity+Pronoun+Inversion+Age+Sex+Region+Social, data = WAS, family=binomial, glmer.control = glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=2e5)))

variables. The GLMM tree shows that, out of the eight predictor variables, seven predictors are chosen as statistically significant at the level of $p < 0.05$; only POLARITY is not selected as significant for the choice of *was* over *were* (see Appendix 2 for the summary of the model). The most important predictor variable, shown in Node 1 at the top, splits the two corpora from one another indicating profound differences in the occurrence of *was* with the pronouns *they*, *we*, and *you* in the 1990s and the 2010s. Overall, looking at the nodes at the bottom of the figure, we see that the highest predicted probabilities for *was* are found under the left-hand branch of the tree, that is BNC1994DS. Reading the glmertree from the top to the bottom, Node 11 has the highest predicted probability of *was*, indicating that within the speakers included in the BNC1994DS, those that belong to the working class, are 30–49 years of age, and live in the South are very highly likely to choose *was* over *were* with pronouns *we* and *you*. Node 10 shows that with the pronoun *they*, male speakers are very likely to use *was*, while for female speakers the probability is clearly lower (cf. Node 9).

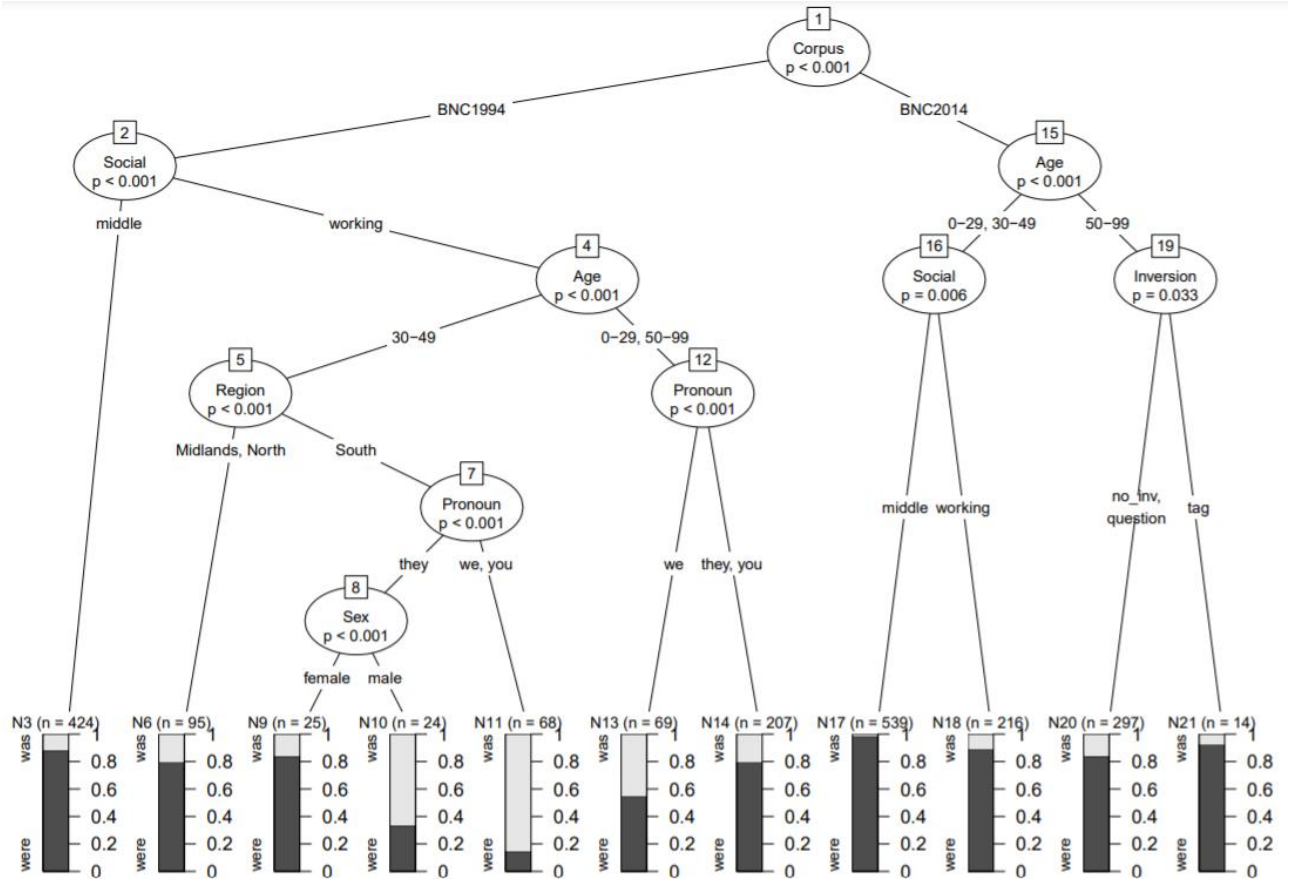

Figure 4.8. GLMM tree predicting the use of *was* with the pronouns *they*, *we* and *you* in BNC1994DS and BNCSpoken2014.

In BNC1994DS, the most important variable is SOCIAL class, separating middle-class speakers as those with the lowest predicted probability of *was*, without any further interactions with other variables. Within the working-class speakers, the oldest and the youngest speakers are grouped together, and their choice of *was* is further governed by an intra-linguistic variable, PRONOUN, so that *we was* is more likely to occur than *you was* or *they was*. Within the middle-aged speakers, REGION makes a further split so that speakers in the Midlands and the North are less likely to use non-standard *was* than speakers from the South. In the South, *was* is the default choice when the subject is *we* or *you*, or with male speakers also *they*. Overall, the data from BNC1994DS seems to conform to findings from earlier research, painting a picture of sociolinguistic variation, based on social class, age, region, and sex, interspersed with variation based on pronoun type.

The newer dataset portrays a much simpler picture in which only the speakers' age and social status, as well as type of inversion, govern the non-standard use of *was*. It is clearly a feature used by the older speakers, of 50 years of age or older, and within the younger age groups, we find a split based on social status, so that *was*-levelling is restricted to the working class (cf. Nodes 17 and 18). Within the older speakers, *was* is less likely to occur in tags than questions or phrases with no inversion (cf. Node 20 and 21). The GLMM tree thus tells us a story of decline, of a move towards standard use – the use of *was* with they/*we*/*you* pronoun subjects has diminished in number and its envelope of variation has tightened (see Section 4.1), and in very recent BrE, *was*-levelling is mostly found in the speech of the older generation and, within the younger speakers, in the lower social classes. On the other hand, we may be dealing with a methodological issue where the small number of tokens in BNCSpoken2014 (*was* RF=79) shrouds the internal variation at a more fine-grained level. Admittedly, the small number of tokens of *they*/*we*/*you was* in BNCSpoken2014 is itself telling of a feature on the decline.

Finally, turning to the variable SPEAKER, which was defined as the random variable in the analysis, Figure 4.9 indicates that some individual speakers portray higher than average probability of choosing *was* (trending towards negative values), while others are more likely to choose *were* (trending towards positive values). A closer look at those individuals showing the most extreme values (above 1.5 or below -1.5) reveals that there are only three individuals with higher probability of *were* (all in BNC1994DS), as opposed to 41 with higher probability of *was* (21 in BNC1994DS, 20 in BNCSpoken2014). Despite the overall trend of working-class or Southern speakers in the BNC1994DS and working-class and older speakers in the BNCSpoken2014 being more likely to choose *was*, a number of individuals counteracting these trends are singled out by the analysis. For

instance, middle-class speakers are reported to favour *was* in both corpora despite the tree in Figure 4.8 clearly showing that, overall, the preference in this speaker group is for *were*. These findings show how the use of *was* with *they*, *we*, and *you* is also subject to intra-speaker variation, in addition to the intra-linguistic and sociolinguistic parameters investigated with the fixed effects.

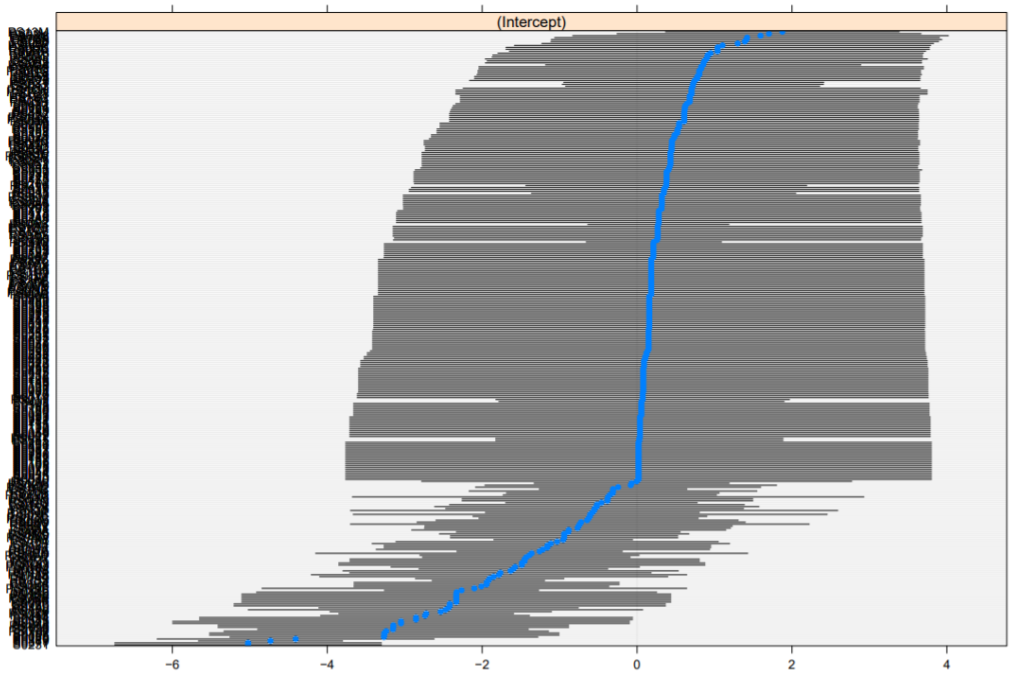

Figure 4.9. Random effects for the GLMM tree in Figure 4.8.

5. Discussion

This study set out to investigate the *was* vs. *were* alternation with the pronouns *they*, *we*, and *you* in recent British English, with the help of the two BNC corpora. Having extracted a dataset of all occurrences of *was* and *were* from both corpora, we annotated the data for intra-linguistic and sociolinguistic variables in order to chart the use of *was* with plural pronouns in recent BrE and to determine the predictor variables that affect the choice of *was* over *were*. Specifically, we asked the following research questions: (i) what diachronic trends can be observed with regard to the *was* vs. *were* alternation with the examined pronoun subjects, (ii) to what extent do age, gender, social class, and dialect area govern the speaker's choice of *was* over *were*, and (iii) to what extent do the pronoun, inversion, or polarity govern the choice?

With regard to research question (i), we found a statistically significant decline in the frequency of *was* with the examined pronouns in BrE (BNC1994DS 7.2%, BNCSpoken2014 1.1%, LL 234.51). At the same time, the envelope of variation has diminished, curtailing *was* to considerably simpler

syntactic contexts in BNCSpoken2014 as opposed to BNC1994DS, as well as to speakers of more specific sociolinguistic backgrounds, so that eventually in BNCSpoken2014, *they*/*we*/*you was* is mostly found in the speech of those of 50 years of age or older, in the North, of males, or of working-class people. It should be noted, however, that the dataset from BNCSpoken2014 in particular is rather small; that is, while we see signs of simplification (movement towards syntactically simpler contexts) and standardization (movement towards standard uses, i.e., *were* with the subject pronouns *they*/*we*/*you*), they remain tentative in nature. Furthermore, the comparability of the two BNC corpora may play a role; Axelsson (2018) points out that BNC2014 consists of more focused conversations, compared to the interactions in BNC1994, thus making it potentially more formal in style. From our perspective, then, this higher level of style potentially leads to less frequent use of *was*-levelling.

The responses to research questions (ii) and (iii) were gleaned with the help of a generalized linear mixed methods tree analysis. In essence, the predictor variables affecting the choice of *was* over *were* are different for the two datasets; in BNC1994DS, the important variables are SOCIAL class, AGE, REGION, PRONOUN, and SEX, whereas in BNCSpoken2014, they are AGE, SOCIAL class, and INVERSION. In general, *was* is more likely to occur in the speech of working-class rather than middle-class speakers, middle-aged rather than the youngest or oldest age groups, and in the North as opposed to the Midlands and the South. In the 1990s, the highest probabilities of *was* are (i) with the pronouns *we* and *you*, in the South, by speakers between the ages of 30 to 49, and in the working class, and (ii) with the pronoun *they* uttered by male speakers of similar age, region, and social class. Within the other age groups, *we was* turns out to be more likely than *you was* or *they was*; these findings conform to Chambers's (2004) and Tagliamonte's (1998) clines of increasing likelihood of *was*, indicating that *they was* is less likely than *you was* or *we was*. However, the present results cannot confirm the difference that polarity has on the choice of *was* over *were*; Tagliamonte (1998) and Anderwald (2002) reported that *was*-levelling is preferred in positive contexts whereas *weren't* is the preferred form in negative contexts, as opposed to *wasn't*. The GLMM tree analysis did not select POLARITY as a significant variable, which may, in fact, be highly reflective of the fact that because of the setup of our study, the analysis measures the degree of non-standard *was* rather than that of non-standard *were* (where the role of polarity has been observed to be of some importance).[8]

[8] A peculiar feature seen in the data relating to non-standard uses of *was*/*were* in tags (with usually reversed polarity) involved the switch between *was* and *were* in the main clause and in the question tag, as in the following instances:

In the most recent dataset, portraying BrE in the 2010s, we witness a loss within the sociolinguistic parameters, as well as a change of intra-linguistic parameters, crucial to the choice of *was* over *were* in recent BrE. The sociolinguistic parameters governing the choice of *was* in the more recent dataset are AGE and SOCIAL class, so that *was* is more likely to occur in the speech of the oldest speakers, or those belonging to the working class within speakers under the age of 50. Within the oldest speakers, *was* is slightly less likely to occur in tags than in questions or contexts without inversion.

Overall, the sharp decrease in the number of occurrences of *was* with the subject pronouns *we*/*you*/*they* between the BNC1994DS and BNCSpoken2014 data is unquestionably drastic, but its characterization in terms of *was*/*were* variation on a broader scale deserves closer attention. The change could easily be seen as suggesting a broad-level shift towards the use of standard patterns with plural pronoun subjects across the UK. This is in line with some findings made in previous studies focussing on specific dialect areas. As noted by Cheshire and Fox (2009, 3), the decline in the use of non-standard *was* has been previously observed in previous studies of larger cities, particularly among younger speakers (e.g., in Birmingham and York, as found in the studies by Khan (2006) and Tagliamonte (1998), respectively). Chambers (2004, 136–139) likewise has argued that urbanization in general is behind the spread of standard English patterns, and in the case of *was*/*were* variation, is a counterforce to the basic tendency towards the use of invariant *was*. Movement towards standard patterns in past tense uses of *be* has also been attested outside of densely populated urban areas, of which Durham's (2013) study on *was*/*were* variation in Shetland English is one example, with evidence of interspeaker variability among younger speakers being seen as indicators of future shifts in favour of standard uses.

As the present study focussed on the pronoun subjects *we*, *you*, and *they*, it is also important to observe the kind of generalizations that the results do not necessarily give grounds for. First of all, it needs to be remembered that the results comment on the degrees of *was*-levelling and not *were*-

(i) So **you were** lucky to be in today really **wasn't you**? (BNC1994DS, KBC)
(ii) so **they was** on a good wage **weren't they**? (BNCSpoken2014, SU8C)

Based on manual inspection of the data, occurrences of the types of (i) and (ii) further indicate that the BNCSpoken2014 data is more adherent to standard uses: BNC1994DS had altogether 12 instances of this type of switch between *was* and *were* between the main clause and the tag (there were altogether 147 relevant tag questions with *was*/*were* + *they*/*we*/*you* in the 1990s data), but the 2010s corpus had only three corresponding cases (out of a total of 213 relevant tag questions). Interestingly enough, all of the speakers of three relevant cases in BNCSpoken2014 were over 60 years of age, whereas similar cases in BNC1994DS were produced by speakers of a wider variety of age groups.

levelling, and the combination of the two in varying degrees can differentiate between dialects, as noted by Schreier (2002). From this point of view observations on shifts towards standard patterns are admittedly partial. On the other hand, *was*-levelling has been argued in general to be more widespread than *were*-levelling, and the degrees of non-standard use of *was* with a specific type of subject offers insights to the study of variation, particularly considering the overall direction towards standard forms in different dialect areas.

As one of the general observations that we can make on the basis of the results of the present work, the BNCSpoken2014 corpus provides plentiful material for variationist linguistic study, particularly when examined in conjunction with the original *British National Corpus*. With a time difference of some 20–25 years between the two sets of data, *was*/*were* variation among British speakers has undergone changes which in some respect are quite drastic, at least as far as the pronoun subjects *we*, *you*, and *they* are concerned. The application of the GLMM tree analysis allows closer inspection of the comparative weight of different intra-linguistic and sociolinguistic parameters as factors relating to the use of standard vs. non-standard patterns. With the pronoun subjects examined, the findings suggest a notable shift towards standard use of *were* in the 2010s, with age (in particular the older speaker groups), region (speakers in Northern England), sex (male speakers), and social class (working class speakers, particularly among the younger speaker groups) standing out as the parameters most strongly associated with the use of non-standard *was*. To complement these findings, further studies into other types of subjects will lead to more nuanced assessments on the variety of dialectal patterns, as well as the broader question of standardization.

Reference list

Anderwald, Lieselotte. 2001. "*Was/were* Variation in Non-standard British English Today." *English World-Wide* 22: 1–22. https://doi.org/10.1075/eww.22.1.02and

Anderwald, Lieselotte. 2002. *Negation in Non-standard British English: Gaps, Regularizations and Asymmetrics*. London: Routledge.

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Axelsson, Karin. 2018. “Canonical Tag Questions in Contemporary British English.” In *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*, edited by Vaclav Brezina, Robbie Love, and Karin Aijmer, 96–119. New York: Routledge.

Britain, David. 2002. “Diffusion, Levelling, Simplification and Reallocation in Past Tense BE in the English Fens.” *Journal of Sociolinguistics* 6: 16–43. https://doi.org/10.1111/1467-9481.00175

Chambers, J. K. 1995. *Sociolinguistic Theory*. Oxford: Blackwell.

Chambers, J. K. 2004. “Dynamic Typology and Vernacular Universals.” In *Dialectology Meets Typology: Dialect Grammar from a Cross-linguistic Perspective*, edited by Bernd Kortmann, 128–145. Berlin: De Gruyter.

Cheshire, Jenny. 1982. *Variation in an English Dialect: A Sociolinguistic Study*. Cambridge: Cambridge University Press.

Cheshire, Jenny and Sue Fox. 2009. “*Was/were* Variation: A Perspective from London.” *Language Variation and Change* 21: 1–38.

Durham, Mercedes. 2013. “*Was/were* Alternation in Shetland English.” *World Englishes* 32 (1): 108–128. http://dx.doi.org/10.1111/weng.12009

Feagin, Crawford L. 1979. *Variation and Change in Alabama English: A Sociolinguistic Study of the White Community*. Washington, D.C.: Georgetown Univ. Press.

Fokkema, Marjolein, Niels Smits, Achim Zeileis, Torsten Hothorn, and Henk Kelderman. 2018. “Detecting Treatment Subgroup Interactions in Clustered Data with Generalized Linear Mixed-effects Model Trees.” *Behavior Research Methods* 50: 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Fokkema, Marjoelin, Julian Edbrooke-Childs, and Miranda Wolpert. 2020. “Generalized Linear Mixed-model (glmm) Trees: A Flexible Decision-tree Method for Multilevel and Longitudinal Data.” *Psychotherapy Research* 31 (22): 1–13. https://doi.org/10.1080/10503307.2020.1785037

Hardie, Andrew. 2012. “CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool.” *International Journal of Corpus Linguistics* 17 (3): 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Hay, Jennifer and Daniel Schreier. 2004. “Reversing the Trajectory of Language Change: Subject–Verb Agreement with *be* in New Zealand English.” *Language Variation and Change* 16 (3): 209–235. https://doi.org/10.1017/S0954394504163047

Khan, Arfaan. 2006. *A Sociolinguistic Study of Birmingham English: Language Variation and Change in a Multi-ethnic British Community.* Unpublished Ph.D. dissertation, University of Lancaster.

Klemola, Juhani. 2000. “The Origins of the Northern Subject Rule: A Case of Early Contact?” In *Celtic Englishes II*, edited by Hildegard Tristram, 329–346. Heidelberg: Winter.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. “The Spoken BNC2014.” *International Journal of Corpus Linguistics* 22 (3): 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Nevalainen, Terttu. 2006. “Vernacular Universals? The Case of Plural *was* in Early Modern English.” In *Types of Variation: Diachronic, Dialectal and Typological Interfaces*, edited by Terttu Nevalainen, Juhani Klemola, and Mikko Laitinen, 351–369. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.76.19nev

RStudio Team. 2021. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Schreier, Daniel. 2002. “Past BE in Tristan da Cunha: The Rise and Fall of Categoricality in Language Change.” *American Speech* 77 (1): 70–99. https://doi.org/10.1215/00031283-77-1-70

Smith, Jennifer and Sali Tagliamonte. 1998. “‘We were all thegither … I think we was all thegither’: *Was* Regularization in Buckie English.” *World Englishes* 17: 105–126. https://doi.org/10.1111/1467-971X.00086

Säily, Tanja, Victorina González-Díaz, and Jukka Suomela. 2018. “Variation in the Productivity of Adjective Comparison in Present-day English.” In *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*, edited by Vaclav Brezina, Robbie Love, and Karin Aijmer, 159–184. London: Routledge.

Tagliamonte, Sali. 1998. “*Was/were* Variation across the Generations: View from the City of York.” *Language Variation and Change* 10 (2): 153–191. https://doi.org/10.1017/S0954394500001277

Tagliamonte, Sali and R. Harald Baayen. 2012. “Models, Forests, and Trees of York English: *Was/were* Variation as a Case Study for Statistical Practice.” *Language Variation and Change* 24: 135–178. https://doi.org/10.1017/S0954394512000129

Wolfram, Walt and Natalie Schilling-Estes. 2003. “Parallel Development and Alternative Restructuring: The Case of *weren’t* Regularization.” In *Social Dialectology: In Honour of Peter Trudgill*, edited by David Britain and Jenny Cheshire, 131–154. Amsterdam: John Benjamins.

Appendices

Appendix 1. Region in the two corpora.

Dialect	BNC1994DS	BNCSpoken2014
Midlands	Midlands, Humberside, Central Midlands, Northeast Mid., South Mid.	Midlands
North	North, Lancashire, Merseyside, Central Northern England, North-east England, Northern England	North
South	South, East Anglian, Home Counties, London, Lower South-west England, Central England, South-west England, Upper South-west England	South

Appendix 2. Summary of the GLMM tree analysis.

Model formula:

WasWere ~ 1 | Corpus + Polarity + Pronoun + Inversion + Age + Sex + Region + Social

Fitted party:

```
[1] root
|   [2] Corpus in BNC1994
|   |   [3] Social in middle: n = 424
|   |       (Intercept)
|   |         2.647468
|   |   [4] Social in working
|   |   |   [5] Age in 30-49
|   |   |   |   [6] Region in Midlands, North: n = 95
|   |   |   |       (Intercept)
|   |   |   |         1.663502
|   |   |   |   [7] Region in South
|   |   |   |   |   [8] Pronoun in they
|   |   |   |   |   |   [9] Sex in female: n = 25
|   |   |   |   |   |       (Intercept)
|   |   |   |   |   |         1.700832
|   |   |   |   |   |   [10] Sex in male: n = 24
|   |   |   |   |   |       (Intercept)
|   |   |   |   |   |         -1.19941
|   |   |   |   |   [11] Pronoun in we, you: n = 68
|   |   |   |   |       (Intercept)
|   |   |   |   |         -1.74338
|   |   |   [12] Age in 0-29, 50-99
|   |   |   |   [13] Pronoun in we: n = 69
|   |   |   |       (Intercept)
|   |   |   |         0.7826827
|   |   |   |   [14] Pronoun in they, you: n = 207
|   |   |   |       (Intercept)
|   |   |   |         2.222229
|   [15] Corpus in BNC2014
```

| | [16] Age in 0-29, 30-49
| | | [17] Social in middle: n = 539
| | | (Intercept)
| | | 4.709034
| | | [18] Social in working: n = 216
| | | (Intercept)
| | | 3.534873
| | [19] Age in 50-99
| | | [20] Inversion in no_inv, question: n = 297
| | | (Intercept)
| | | 2.784401
| | | [21] Inversion in tag: n = 14
| | | (Intercept)
| | | 5.48883

Number of inner nodes: 10

Number of terminal nodes: 11

Number of parameters per node: 1

Objective function (negative log-likelihood): 404.6728