The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: feature extraction, coding protocols, projections to geography, summary statistics

Benedikt Szmrecsanyi
Freiburg Institute for Advanced Studies
<bszm@frias.uni-freiburg.de>

Recent and forthcoming Freiburg-based dialectometrical research investigates aggregate morphosyntactic variability in dozens of traditional British English dialects drawing on a measure of joint dialectal distance. This measure is calculated on the basis of a catalogue spanning 57 morphosyntax features (e.g. WOULD as marker of habitual past, *never* as past tense negator, *don't* with 3rd person singular subjects, and so on) whose text frequencies were established through querying the *Freiburg Corpus of English Dialects*. This document details the technicalities of the feature extraction process, and the coding protocols underpinning it. The discussion also encompasses cartographic projections of feature frequency distributions, as well as some summary statistics.

Contents

1. Introduction

This document is concerned with the technicalities, on a feature-by-feature basis, behind the feature catalogue that forms the empirical basis for the aggregative-dialectometrical research in Szmrecsanyi (2010; submitted a,b; in preparation). Most of the features discussed here also feed into the analysis presented in Szmrecsanyi (2008).

2. Data source: the *Freiburg Corpus of English Dialects*

We detail how the features were coded in and extracted from the *Freiburg Corpus of English Dialects* (FRED) (see Hernández 2006; Szmrecsanyi and Hernández 2007 for manuals, and `http://www.helsinki.fi/varieng/CoRD/corpora/FRED/` for general information). The version of the corpus drawn upon contains 368 individual texts and spans approximately 2.44 million words of running text, consisting of samples (mainly transcribed so-called 'oral history' material) of dialectal speech from a variety of sources. Most of these samples were recorded in the 1970's and 1980's. Usually, a fieldworker interviews an informant about life, work etc. in former days. The 427 informants sampled in the corpus are typically elderly males with a working-class background (so-called NORMS). The interviews were conducted in 156 different locations (that is, villages and towns) in 34 different pre-1974 counties in Great Britain including the Isle of Man and the Hebrides (specifically: Cornwall, Devon, Somerset, Kent, Wiltshire, London, Middlesex, Glamorganshire, Oxfordshire, Suffolk, Warwickshire, Shropshire, Leicestershire, Nottinghamshire, Denbighshire, Lancashire, Isle of Man, Yorkshire, Westmorland, Durham, Dumfriesshire, Northumberland, Selkirkshire, Peebleshire, Midlothian, East Lothian, West Lothian, Perthshire, Angus, Kincardineshire, Hebrides, Banffshire, Ross and Cromarty, Sutherland). FRED is at present neither syntax-parsed nor part-of-speech annotated.

3. The feature catalogue: an overview

What follows is a list of features in the catalogue, annotated with linguistic examples. The features are typically well-known phenomena amply discussed in the dialectological, variationist, and corpus-linguistic literature. The catalogue – which overlaps with, but is not identical to, the list of phenomena investigated in the comparative morphosyntax survey by Kortmann and Szmrecsanyi (2004) and the battery of morphosyntax features covered in *Survey of English Dialects*-based interpretation atlases (cf. Orton et al. 1978; Viereck et al. 1991) – spans eleven major grammatical domains:

A. Pronouns and determiners

[1] non-standard reflexives (e.g. *they didn't go* theirself)
[2] standard reflexives (e.g. *they didn't go* themselves)
[3] archaic *thee/thou/thy* (e.g. *I tell* thee *a bit more*)
[4] archaic *ye* (e.g. ye*'d dancing every week*)
[5] *us* (e.g. us *couldn't get back, there was no train*)
[6] *them* (e.g. *I wonder if they'd do any of* them things *today*)

B. The noun phrase

[7] synthetic adjective comparison (e.g. *he was always* keener *on farming*)
[8] the *of*-genitive (e.g. *the presence* of *my father*)
[9] the *s*-genitive (e.g. *my father*'s *presence*)
[10] preposition stranding (e.g. *the very house* which *it was* in)
[11] cardinal number + *years* (e.g. *I was there about* three years)
[12] cardinal number + *year*-Ø (e.g. *she were* three year *old*)

C. Primary verbs

[13] the primary verb TO DO (e.g. *why* did *you not wait?*)
[14] the primary verb TO BE (e.g. *I* was *took straight into this pitting job*)
[15] the primary verb TO HAVE (e.g. *we thought somebody* had *brought them*)
[16] marking of possession – HAVE GOT (e.g. *I* have got *the photographs*)

D. Tense and aspect

[17] the future marker BE GOING TO (e.g. *I*'m going to *let you into a secret*)
[18] the future markers WILL/SHALL (e.g. *I* will *let you into a secret*)
[19] WOULD as marker of habitual past (e.g. *he* would *go around killing pigs*)
[20] *used to* as marker of habitual past (e.g. *he* used to *go around killing pigs*)
[21] progressive verb forms (e.g. *the rest* are going *to Portree School*)
[22] the present perfect with auxiliary BE (e.g. *I*'m come *down to pay the rent*)
[23] the present perfect with auxiliary HAVE (e.g. *they*'ve killed *the skipper*)

E. Modality

[24] marking of epistemic and deontic modality: MUST (e.g. *I* must *pick up the book*)
[25] marking of epistemic and deontic modality: HAVE TO (e.g. *I* have to *pick up the book*)
[26] marking of epistemic and deontic modality: GOT TO (e.g. *I* gotta *pick up the book*)

F. Verb morphology

[27] *a*-prefixing on *-ing*-forms (e.g. *he was* a-waiting)
[28] non-standard weak past tense and past participle forms (e.g. *they* knowed *all about these things*)
[29] non-standard past tense *done* (e.g. *you came home and* done *the home fishing*)
[30] non-standard past tense *come* (e.g. *he* come *down the road one day*)

G. Negation

[31] the negative suffix *-nae* (e.g. *I can*nae *do it*)
[32] the negator *ain't* (e.g. *people* ain't *got no money*)
[33] multiple negation (e.g. *do*n't *you make* no *damn mistake*)
[34] negative contraction (e.g. *they* won't *do anything*)
[35] auxiliary contraction (e.g. *they*'ll not *do anything*)
[36] *never* as past tense negator (e.g. *and they* never *moved no more*)
[37] WASN'T (e.g. *they* wasn't *hungry*)
[38] WEREN'T (e.g. *they* weren't *hungry*)

H. Agreement

[39] non-standard verbal *-s* (e.g. *so* I says, *What have you to do?*)
[40] *don't* with 3rd person singular subjects (e.g. *if this man* don't *come up to it*)
[41] standard *doesn't* with 3rd person singular subjects
(e.g. *if this man* doesn't *come up to it*)
[42] existential/presentational *there is/was* with plural subjects
(e.g. there was children *involved*)
[43] absence of auxiliary BE in progressive constructions
(e.g. *I said, How* Ø you doing*?*)
[44] non-standard WAS (e.g. *three of them* was *killed*)
[45] non-standard WERE (e.g. *he* were *a young lad*)

I. Relativization

[46] *wh*-relativization (e.g. *the man* who *read the book*)
[47] the relative particle *what* (e.g. *the man* what *read the book*)
[48] the relative particle *that* (e.g. *the man* that *read the book*)

J. Complementation

[49] *as what* or *than what* in comparative clauses
(e.g. *we done no more* than what *other kids used to do*)
[50] unsplit *for to* (e.g. *it was ready* for to *go away with the order*)
[51] infinitival complementation after BEGIN, START, CONTINUE, HATE, and LOVE
(e.g. *I* began to take *an interest*)
[52] gerundial complementation after BEGIN, START, CONTINUE, HATE, and LOVE
(e.g. *I* began taking *an interest*)
[53] zero complementation after THINK, SAY, and KNOW
(e.g. *they just thought* [Ø *it isn't for girls*])
[54] *that* complementation after THINK, SAY, and KNOW
(e.g. *they just thought* [that *it isn't for girls*])

K. Word order and discourse phenomena

[55] lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes/no*-questions (e.g. *where* Ø *you put the shovel?*)
[56] the prepositional dative after the verb GIVE (e.g. *she gave* [*a job*] [to *my brother*])
[57] double object structures after the verb GIVE (e.g. *she gave* [*my brother*] [*a job*])

4. Some general remarks on the feature extraction process

In terms of the technicalities of the extraction process, the items in the feature catalogue fall into two broad groups: features that could be extracted fully automatically, as opposed to features that required manual coding prior to the actual extraction procedure.

4.1. Fully automatic extraction

31 features are sufficiently 'surfacy' to be extractable without human intervention. For instance, feature [32] (*ain't*) (cf. Section 5.21) can be reliably identified by computer software. In such cases, a retrieval script written in the programming language *Perl* (cf. `http://www.perl.org/`) identified occurrences of the target phenomenon in the dataset and established the relevant text frequencies per FRED county and location.

4.2. Semi-automatic extraction

26 features in the catalogue require manual disambiguation prior to the actual extraction procedure, and in all the dataset as a whole owes its existence to a total of well over 80,000 manual coding decisions. For instance, feature [48] (the relative particle *that*) (cf. Section 5.32) cannot be automatically extracted by software as the untagged, non-parsed corpus material in FRED does not permit reliable automatic disambiguation between relativizer *that* and other uses of *that* (such as demonstrative *that* or complementizer *that*). In cases like this, we typically relied, first, on screening scripts written in *Perl* to considerably narrow down the number of phenomena which had to be inspected manually. For example, in the case of relativizer *that*, a screening script discarded the sequence *when that*, in which *that* can never function as a relativizer. After the pre-screening process, the remaining corpus hits were inspected manually and annotated if they constituted instantiations of the target phenomenon. In a last step, retrieval scripts relied on the manual annotation to automatically establish the relevant text frequencies per FRED county and location.

The manual coding process usually involved the full FRED dataset. However, in the case of nine features ([19], [20], [21], [23], [44], [45], [46], [47], and [48]) with fairly high text frequencies, we economized by manually inspecting not the full dataset, but an abridged version spanning 510,000 words of running text. It contains the same texts (interviews) as the full FRED dataset, the difference being that in case of texts longer than 1,500 words of running text, only the first 1,500 words were sampled into the abridged dataset (all other texts are represented in full length). To illustrate: Feature [19] (WOULD as a marker of habitual past) (cf. Section 5.12) is known to be a rather frequent feature (especially in oral history interviews), yet its identification in textual material also implicates a good deal of manual coding. Thus, we manually inspected $\text{FRED}_{\text{abridged}}$, where the phenomenon has 2,119 occurrences, which corresponds to a mean normalized text frequency of 42 *pttw* across FRED (cf. Table 1, p. 62).

5. Coding protocols for individual features

This section details how occurrences of each of the 57 features in the catalogue were identified (manually, if necessary) in the dataset. For every feature – or, where applicable, every pairing of closely related features – we provide:

- a description of the technicalities of the extraction process, including the coding protocols (where applicable) guiding this process;
- a boxplot that depicts variability by a-priori dialect area (roughly following Trudgill's dialect division on pronunciational grounds [Trudgill 1999: Map 9]), on the basis of frequency variance across the $N = 156$ locations sampled in the corpus;
- a projection to geography that depicts relative text frequency variability on the level of the 34 FRED counties.

A few technical comments on the boxplots and the cartographic projections are in order here. In the boxplots, the boxes depict the interquartile frequency range comprising the middle 50% of all frequency observations, with the thick line in the boxes indicating the median. The whiskers to the left and right of the boxes (where applicable) extend to data points that score no more than 1.5 times the interquartile range. As for the projections to geography, we first utilized customary Voronoi tesselation (Voronoi 1907) to assign every FRED county a convex polygon in geographic map space such that every point within the polygon is closer to the generating dialect site than to any other dialect site; the radius of the Voronoi polygons was limited to approximately 50km in order to do visual justice to the areal coverage of the dialect corpus. Map 1 displays the resulting tesselation grid. Subsequently, the observable frequency range was divided into 10 percentile groups such that each group contained roughly the same number of observations. Each county's 10-tile rank was then mapped on the red-green-blue color scheme, assigning a perfect red hue to the highest 10-tile group, a perfect blue hue to the lowest 10-tile group, and gradient red-blue color blends to the 10-tile groups in between. This is another way of saying that in the maps, more reddish tones indicate relatively higher text frequencies whereas more blueish tones indicate relatively lower text frequencies.

5.1. Reflexive pronouns: feature [1] (non-standard reflexives) / feature [2] (standard reflexives)

All occurrences of non-standard, regularized reflexive pronoun forms (*hisself, theirselves, theirself, ourself*), as in (1), as well as all occurrences of standard reflexive pronoun forms (*myself* [*miself*], *yourself, herself, himself, itself, ourselves, yourselves, themselves*), as in (2), were automatically identified and registered by a retrieval script:

(1) a. ... he filled *hisself* up with cod ... <FRED SFK005>

ANS	Angus
BAN	Banffshire
CON	Cornwall
DEN	Denbighshire
DEV	Devon
DFS	Dumfriesshire
DUR	Durham
ELN	East Lothian
GLA	Glamorganshire
HEB	Hebrides
MAN	Isle of Man
KCD	Kincardineshire
KEN	Kent
LAN	Lancashire
LEI	Leicestershire
LND	London
MDX	Middlesex
MLN	Midlothian
NBL	Northumberland
NTT	Nottinghamshire
OXF	Oxfordshire
PEE	Peebleshire
PER	Perthshire
ROC	Ross and Cromarty
SAL	Shropshire
SEL	Selkirkshire
SFK	Suffolk
SOM	Somerset
SUT	Sutherland
WAR	Warwickshire
WES	Westmorland
WIL	Wiltshire
WLN	West Lothian
YKS	Yorkshire

Map 1: The FRED tesselation grid (level of granularity: $N = 34$ counties).

Figure 1: Feature [1] (non-standard reflexives). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

b. They make *theirselves* fast to a stone or shell or that. <FRED SFK030>
c. they didn't go *theirself*, but they made sure you went ... <FRED LAN003>
d. Yes, we made that *ourself*. <FRED SOM004>

(2) No they just fend for *themselves* ... <FRED HEB038>

5.2. 2nd person singular pronouns: feature [3] (archaic *thee*, *thou*, and *thy*) / feature [4] (archaic *ye*)

All occurrences of the archaic, non-standard 2nd person singular pronouns (*thou, thee*) and possessive determiners (*thy*), as in (3), as well as all occurrences of the form *ye*, as in (4), were automatically identified and registered by a retrieval script.

(3) a. ... I asked, does *thou* wash *thy* hands after *thou* use that stuff? <FRED SAL008>
b. Ah, I tell *thee* a bit more yet. <FRED SOM028>.

(4) Aye, *ye*'d dancing every week at that time. <FRED ANS001>

Notice that the form *thine* only attests 1 occurrence in FRED, and was not considered.

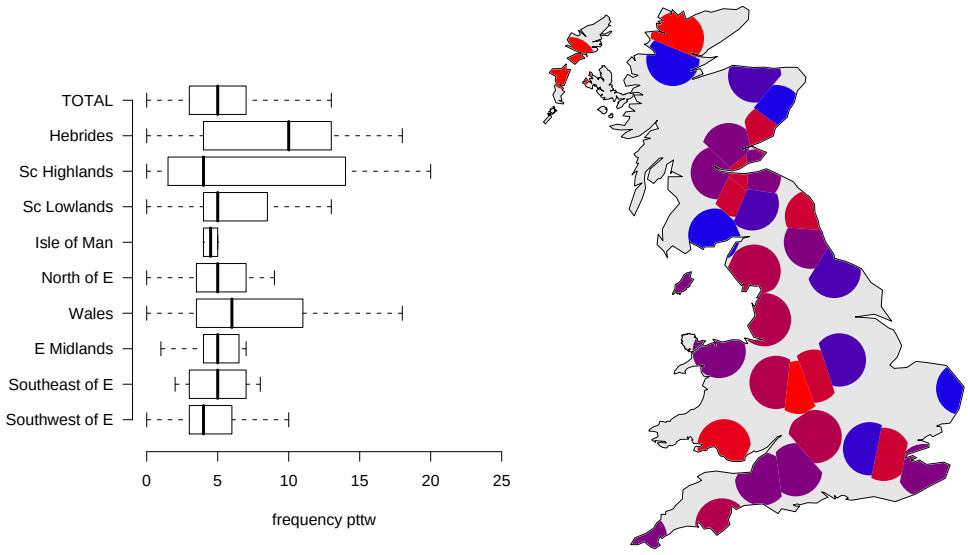

Figure 2: Feature [2] (standard reflexives). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

Figure 3: Feature [3] (archaic *thee*, *thou*, and *thy*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

Figure 4: Feature [4] (archaic *ye*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.3. 1st person plural pronouns and determiners: feature [5] (*us*)

A retrieval script identified and registered all occurrences of the form *us*. These include occurrences of standard *us* (as in (5)), non-standard usage as a subject pronoun (as in (6)), and possibly non-standard usage as a possessive determiner.

(5) ... they wouldn't give *us* sweets and nuts and things ... <FRED HEB010>

(6) *Us* couldn't get back, there was no train. <FRED DEV007>

5.4. Demonstrative determiners: feature [6] (*them*)

A retrieval script identified and registered all occurrences of the form *them* followed by a token ending in *-s* (as in (7)), suggesting a following plural noun. Note that occurrences of *them* followed by the following high-frequency collocates (which are clearly not plural nouns) were ignored: *as, was, this, is, across, sometimes, perhaps, indoors, afterwards, acrorss, themselves, has, always, upstairs, theirselves, downstairs, discuss.*

(7) I wonder if they'd do any of *them things* today? <FRED SAL017>

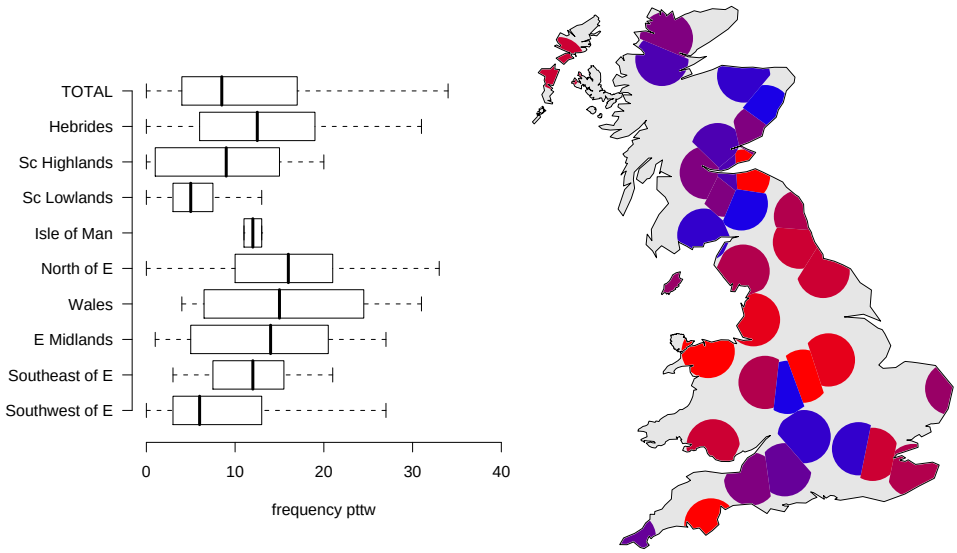

Figure 5: Feature [5] (*us*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

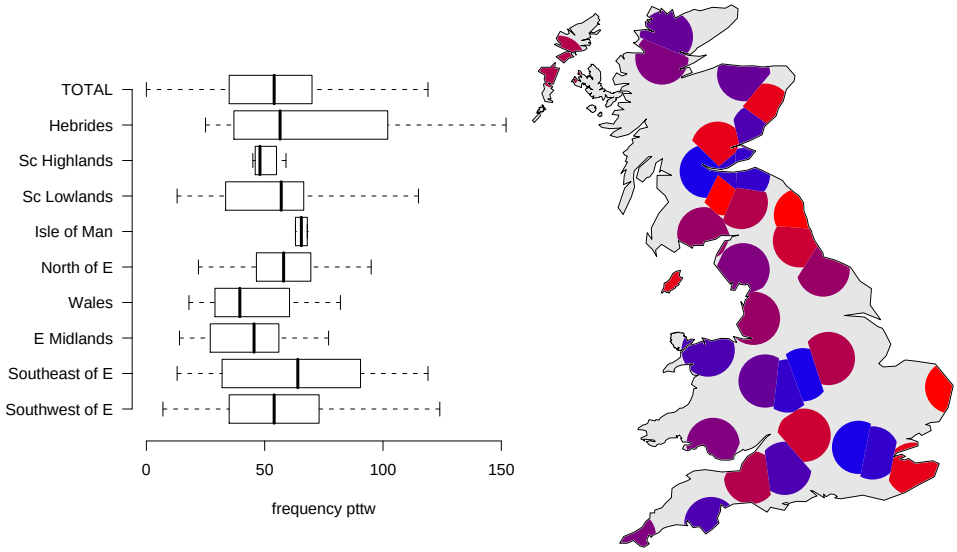

Figure 6: Feature [6] (*them*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

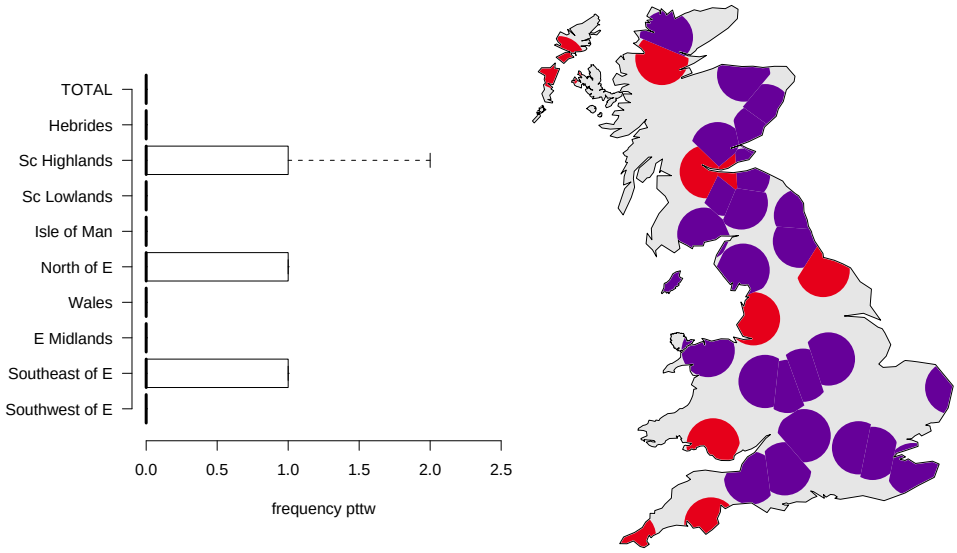

Figure 7: Feature [7] (synthetic adjective comparison). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.5. Adjective comparison: feature [7] (synthetic adjective comparison)

Following the methodology detailed in Szmrecsanyi (2006: 67–69) and Szmrecsanyi (2005), a list of 112 adjectives, which are known to take both synthetic and analytic comparison, was defined (e.g. Bauer 1994: 55; Biber et al. 1999: 522; Leech and Culpeper 1997: 356–364; Mondorf 2003: 257, 287; Quirk et al. 1985: 462): *able, acute, afraid, akin, ample, apt, aware, bitter, bizarre, blunt, bold, brittle, cheap, cheeky, clear, clever, common, compact, complete, correct, costly, cosy, crazy, cruel, curt, dead, deadly, dense, empty, exact, extreme, feeble, fierce, fit, fond, free, friendly, full, gentle, guilty, handsome, handy, humble, hungry, intense, just, keen, kindly, likely, little, lively, lonely, lovely, lowly, lucky, mature, mellow, narrow, nimble, noble, obscure, odd, pale, pleasant, polite, poor, precise, profane, profound, prone, proud, queer, quiet, rare, ready, real, remote, rich, right, risky, robust, rude, secure, severe, sexy, shallow, sick, silly, simple, sincere, slender, slow, sober, solid, sound, stable, stupid, subtle, sure, tender, trendy, tricky, true, ugly, unhappy, unwise, used, wealthy, wicked, worthy, wrong, yellow.*

Subsequently, a retrieval script identified and registered all occurrences of adjectives in the above list which were suffixed by *-er*, as in (8);

(8) He was always *keener* on farming than I was. <FRED CON008>

5.6. Genitives: feature [8] (the *of*-genitive) / feature [9] (the *s*-genitive)

A screening script processed more than 30,000 instances of the token *of* and more than 5,000 instances of the clitic or contraction *'s* in the dataset, weeding out a total of 168 collocational patterns (such as *of course, of age, dozens of, lots of*, etc.) where an interchangeable genitive construction can be ruled out *a priori*. Subsequently, the more than 7,000 remaining instances of *of* and about 3,000 remaining instances of *'s* were inspected manually/qualitatively as to whether they constitute instances of roughly interchangeable genitives, as in (9), or not, as in (10):

(9) a. ... somehow the presence *of* my father seems to make me a bit more embarrassed. (~ my father*'s* presence) <FRED HEB009>
b. ... I kept mi mother*'s* house clean for twenty five year ... (~ the house *of* mi mother) <FRED DUR003>

(10) a. No. I think it was a matter *of* suicide. (?a suicide*'s* matter) <FRED IOM001>
b. ... and then the middle'd come out in half an hour*'s* time. (?in the time *of* half an hour) <FRED SAL016>

Tags were manually inserted when an interchangeable genitive was identified. The coding scheme detailed in Hinrichs and Szmrecsanyi (2007: 444–448) (cf. also Szmrecsanyi and Hinrichs 2008) guided the coding procedure. Thus, a negative list of non-interchangeable types and cases constrained the coder's judgments of interchangeability. For example, the following *s*-genitive types were excluded from analysis:

1. any construction in which a genitive *'s* is not followed by a possessum phrase, as in (11):

   (11) Come on, let's go down to Aunty Agnes*'s*. <FRED NTT012>

2. any idiomatic phrase that has been conventionalized with the s-genitive, as in (12):

   (12) Two journeys a day was a good *day's work*. <FRED SAL033>

Analogously, some *of*-genitive types excluded from the analysis are the following:

1. *of*-genitives with indefinite possessum phrases, as in (13):

   (13) ... he was a secretary *of* some un- union ... <FRED LAN002>

2. measures expressed with *of*-constructions, as in (14):

   (14) I said, Well, I said, I want about two *pound of* nice fat pork. <FRED KEN006>

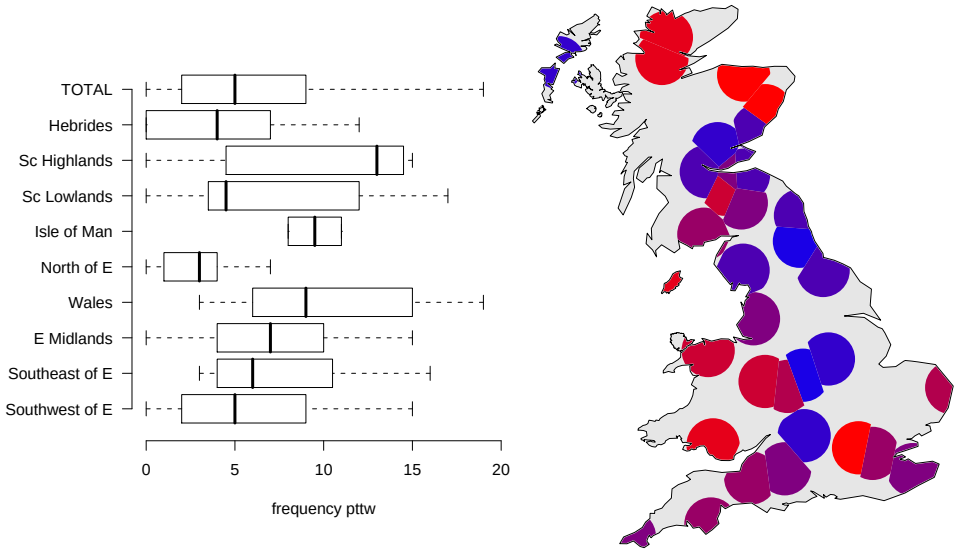

Figure 8: Feature [8] (the *of*-genitive). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

3. any phrase that has been conventionalized with the *of*-genitive, as in (15):

   (15) I remember *the Prince of Wales* ... <FRED NBL006>

In a last step, the manually inserted tags were automatically identified and registered utilizing a retrieval script.

5.7. Preposition stranding: feature [10]

A list of the 10 most frequent prepositions in the *British National Corpus* (cf. Aston and Burnard 1998) was generated: *in, to, on, for, with, at, about, from, by, into* (POS tag: PRF/PRP). Subsequently, a screening script identified all instances of these tokens in the FRED dataset, flagging them for manual analysis if they were likely to be clause-final elements, as suggested by the presence of

- an immediately following punctuation mark (< . , ; ! ? >), as in (16):

  (16) I don't know where he got them *from*, but – <FRED CON003>

- an immediately following conjunction: *and, but, if, or, cos, when, because, as, than, so, before, whether, where, while, unless, after, once, until, like, even, since,*

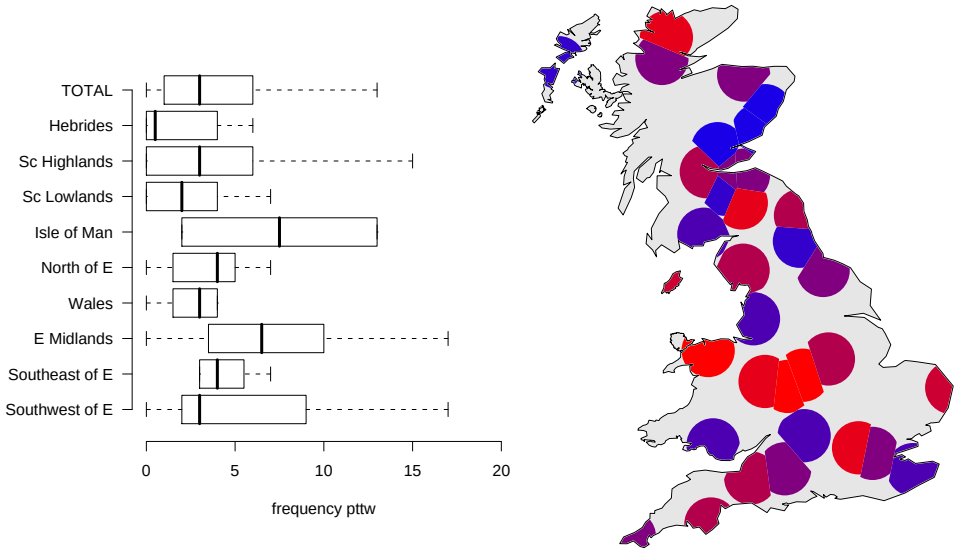

Figure 9: Feature [9] (the *s*-genitive). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

till, although, whereas (these conjunctions are the most frequent ones in the *British National Corpus* [POS tag: CJC/CJS/CJT]), as in (17):

(17) . . . in each corner was the initial of the man who who the property belongs *to and* his wife's initials . . . <FRED DEV004>

At the same time, the screening script ignored a number of contexts that fairly reliably rule out the presence of a stranded preposition:

- contexts strongly suggesting a phrasal/prepositional verb construction, as in (18):

  (18) You used to send in notice when you had a few that would do, and they would *come in* and collect them within a few days. <FRED CON007>

- set phrases, as in (19):

  (19) So that gave me a bit of perks and helped, you know, with, because we retired to this place rather *early on*. <FRED CON011>

This procedure yielded more than 9,000 tokens in the dataset potentially instantiating a stranded preposition. These tokens were inspected manually/qualitatively by a native speaker of English and tagged if a stranded preposition was indeed instantiated. Note

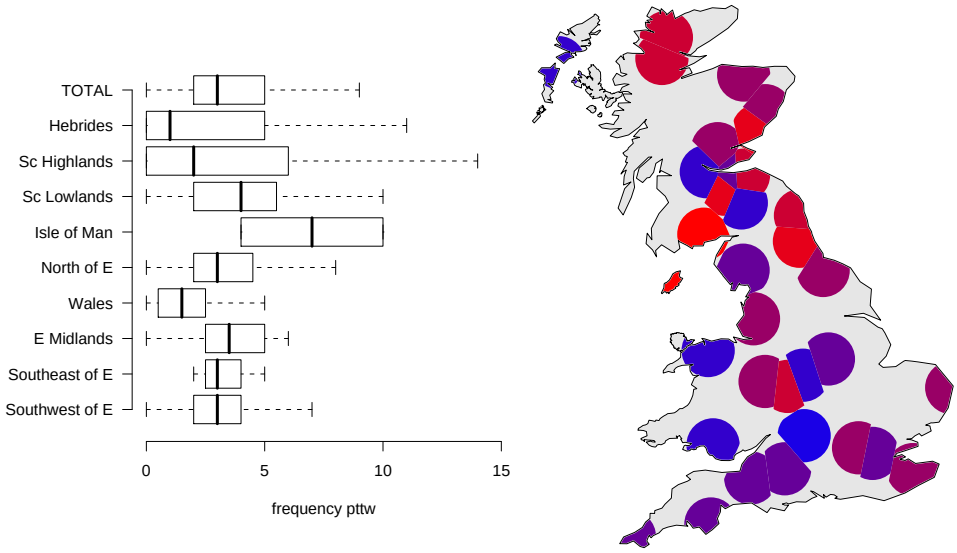

Figure 10: Feature [10] (preposition stranding). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

that tokens were tagged only if they occurred in relative clauses (as in (20-a)), nominal clauses (as in (20-b)), or indirect questions (as in (20-c)):

(20) a. ...and my father was leaving the farm [he was *on*] ... <FRED ANS001>
b. This is [what you sat *on*] ... <FRED ANS001>
c. ...he dare not ask them [where the money went *to*]. <FRED GLA001>

Crucially, a preposition was considered stranded if and only if it could also have occurred in a pied piping construction before a *wh*-element. So, a semantically roughly equivalent syntactic alternative to (20-a) is (21):

(21) ...and my father was leaving the farm [*on* which he was] ...

Finally, a retrieval script identified and registered all stranded prepositions as identified by the manual inspection procedure.

5.8. Plural marking: feature [11] (cardinal number + *years*) / feature [12] (cardinal number + *year*-Ø)

A retrieval script automatically identified and registered all instances of the forms *years* (cf. (22)) or *year* (cf. (23)) preceded by the following tokens: *two*, three*, four*,*

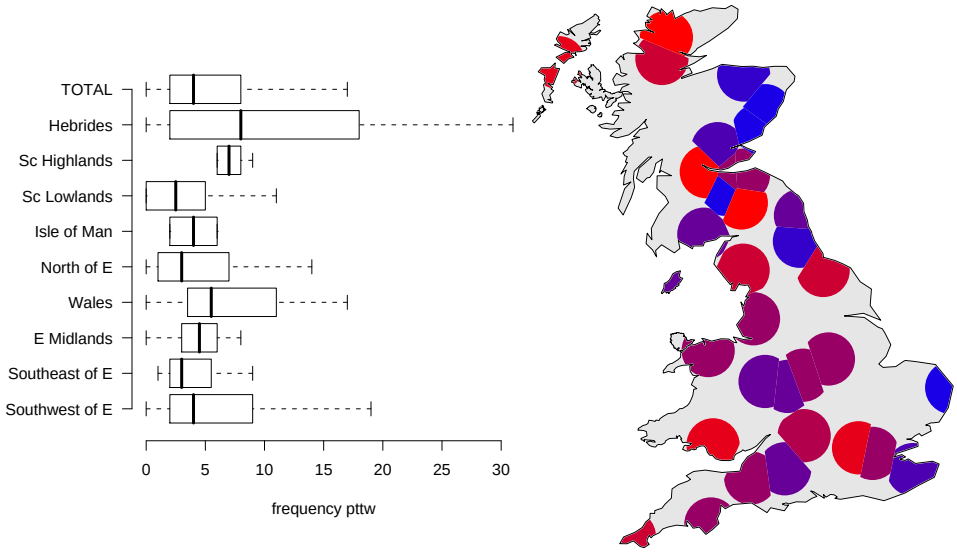

Figure 11: Feature [11] (cardinal number + *years*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

five, six*, seven*, eight*, nine*, ten*, eleven*, twelve*, thirteen*, fourteen*, fifteen*, twenty*, thirty*, fifty*, hundred*, thousand**.[1]

(22) I was there about *three years* I think. <FRED WES014>

(23) She were *three year* old when I got home. <FRED NTT004>

5.9. Primary verbs: feature [13] (TO DO) / feature [14] (TO BE) / feature [15] (TO HAVE)

A retrieval script automatically identified and registered text frequencies of the following forms:

- TO DO – the forms *do, does, did, done, don't, doesn't, doesnae, dussn't, didn't, didnae, dint, din't, doing, doin'*. No distinction was made between auxiliary *do*, as in (24-a), and main verb *do*, as in (24-b).

  (24) a. Why *did* you not wait till about ten or eight o'clock? <FRED HEB018>

1 Here and in the following, the asterisk (*) is used as a wildcard that represents zero or more characters.

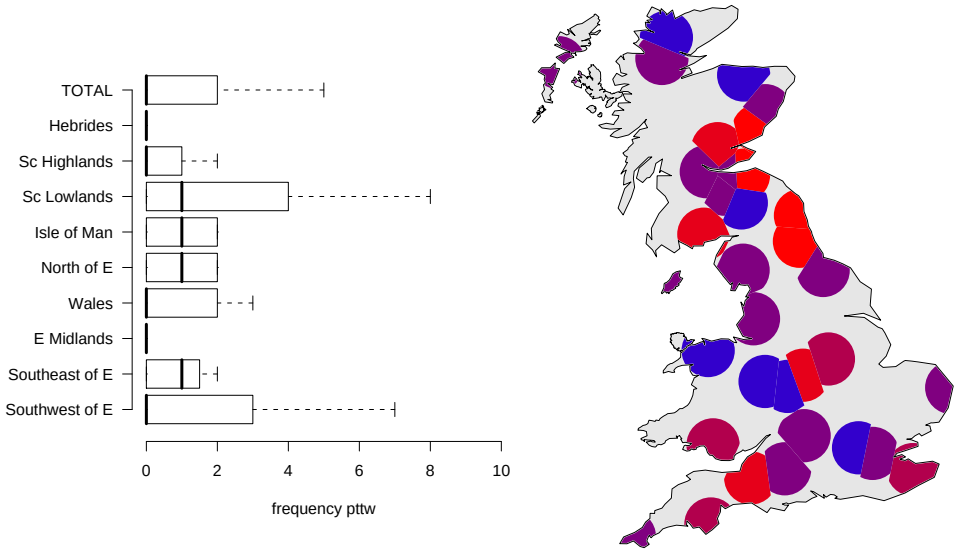

Figure 12: Feature [12] (cardinal number + *year*-Ø). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

b. We *did* no trawling then ... <FRED KEN008>

- TO BE – the forms *be, am, amn't, 'm, is, isn't, isnae, 's, are, 're, was, wasn't, wasnae, were, weren't, werenae, been, being, bein'*. No distinction was made between auxiliary *be*, as in (25-a), and main verb *be*, as in (25-b).

(25) a. ... I *was* took straight into this pitting job ... <FRED LEI002>
b. ... and eh he worked there till he *was* sixty-five ... <FRED LEI002>

- TO HAVE – the forms *have, has, hasn't, hasnae, 've, had, hadn't, hadnae, having, havin'*. No distinction was made between auxiliary *have*, as in (26-a), and main verb *have*, as in (26-b).

(26) a. ... we thought somebody *had* brought them you know ... <FRED NTT005>
b. And they'd been there till then, and they never *had* no annual holiday at all. <FRED NTT005>

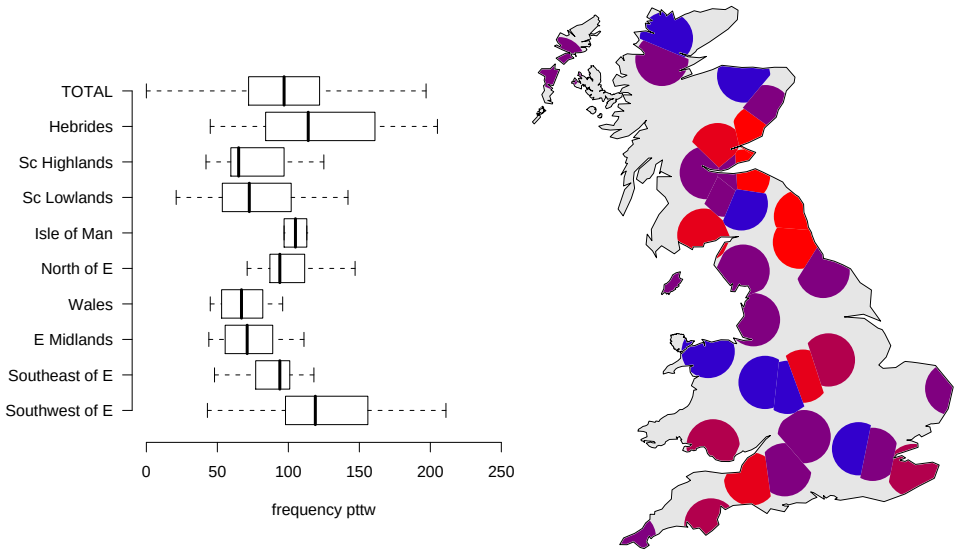

Figure 13: Feature [13] (TO DO). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

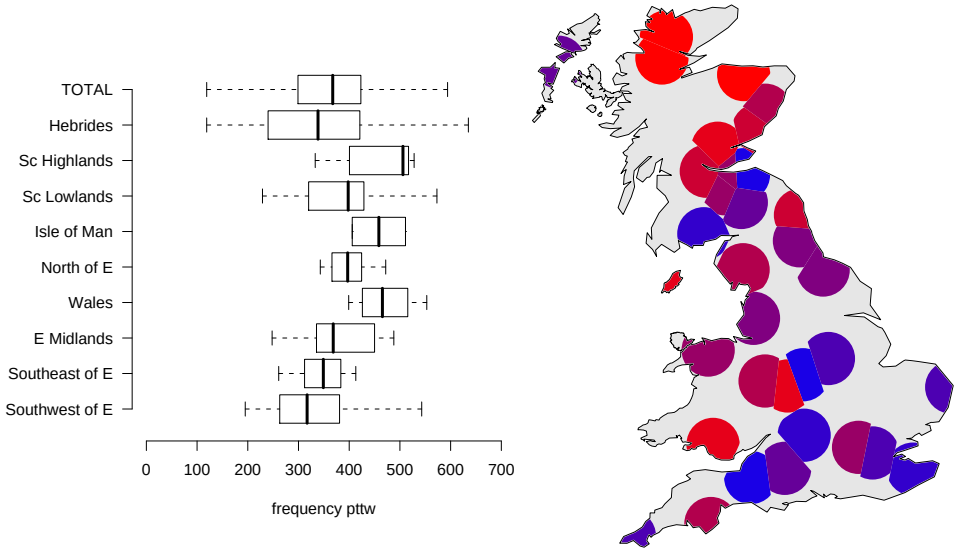

Figure 14: Feature [14] (TO BE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

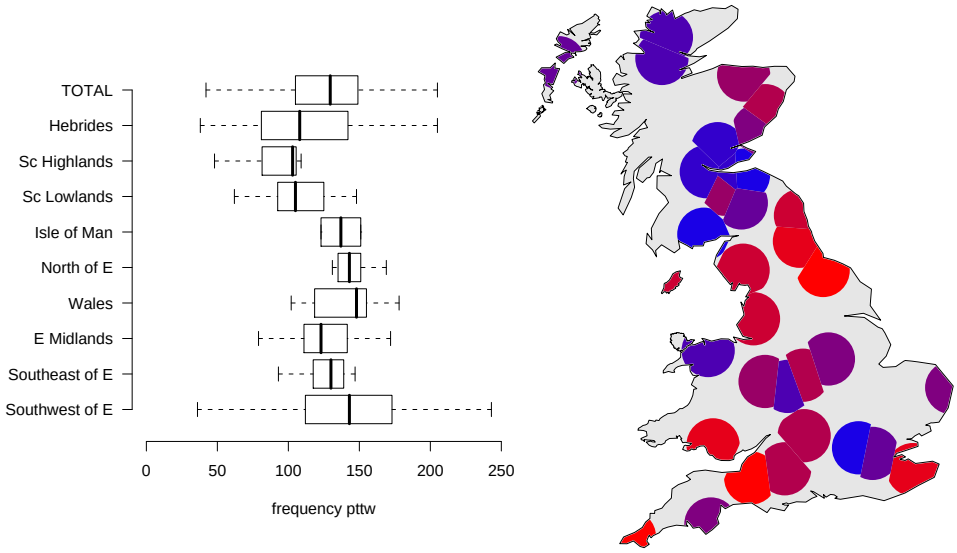

Figure 15: Feature [15] (TO HAVE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.10. Marking of possession: feature [16] (HAVE GOT)

A retrieval script identified and registered all occurrences of the forms *have, has, hasn't, hasnae, 've, had, hadn't, hadnae, having, havin'* followed by the token *got*, as in (27). Instances where the HAVE GOT sequence was followed by the token *to* (as in (28-a)), indicating marking of obligation and not possession, were ignored, as were HAVE GOT sequences followed by the particles *up*, *back*, *in*, and *out* (as in (28-b)):

(27) Yes, I *have got* the photographs. <FRED LAN023>

(28) a. Police *have got to* get a warrant and all that ... <FRED DEV006>
b. ...no doubt I could *have got in* the Albion colliery ... <FRED GLA003>

5.11. Future markers: feature [17] (BE GOING TO) / feature [18] (WILL/SHALL)

About 2,000 instances in the dataset of the forms *going to* and *will* were inspected to determine if the forms function as future markers, in which case a tag was manually inserted. This means, more specifically, that an occurrence of *going to* had to be followed by a bare infinitive as opposed to an NP (as in (29-a)), and that the form *will* had to be a verb, not a noun (as in (29-b)).

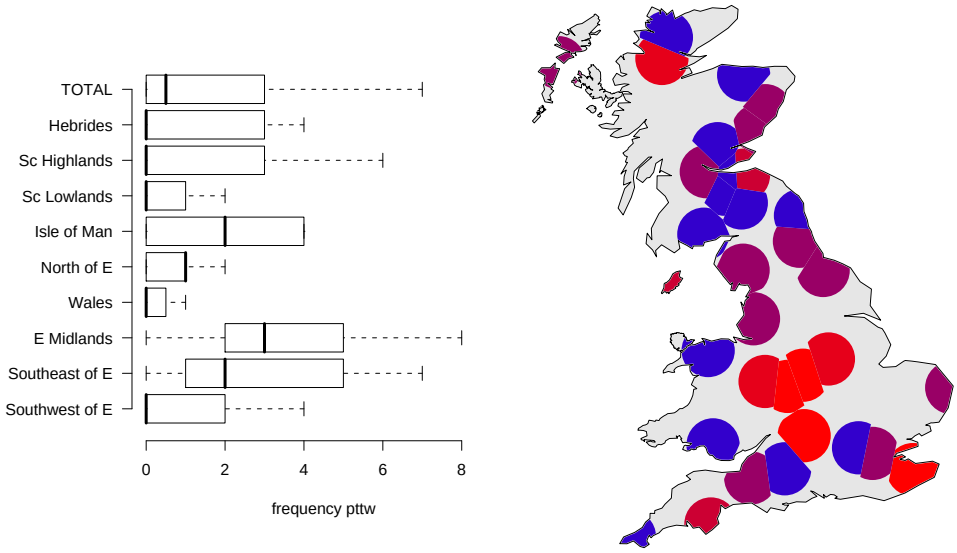

Figure 16: Feature [16] (HAVE GOT). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

(29) a. I remember *going to* church ... <FRED LAN008>
b. Well, actually, she lost the *will* to live ... <FRED KEN008>

A retrieval script then identified and registered

- text frequencies of future marking BE GOING TO (i.e., frequencies of the form *gonna* as well as of manually tagged occurrences of the form *be going to*, as in (30):

  (30) a. ... I*'m going to* let you into a secret. <FRED WES010>
  b. ... Where you *gonna* take those pigs, boy? <FRED KEN002>

- text frequencies of future marking WILL/SHALL, i.e., frequencies of the forms *won't, 'll, shall, shan't* as well as of manually tagged occurrences of the form *will*, as in (31):

  (31) a. I *won't* run after him. <FRED SOM032>
  b. Oh I*'ll* tell you man. <FRED GLA002>
  c. Yes, I *shall* be 81 next month ... <FRED NTT006>
  d. I *shan't* say no names. <FRED SFK009>
  e. ... he says, You *will* be fined fifteen shillings. <FRED KEN003>

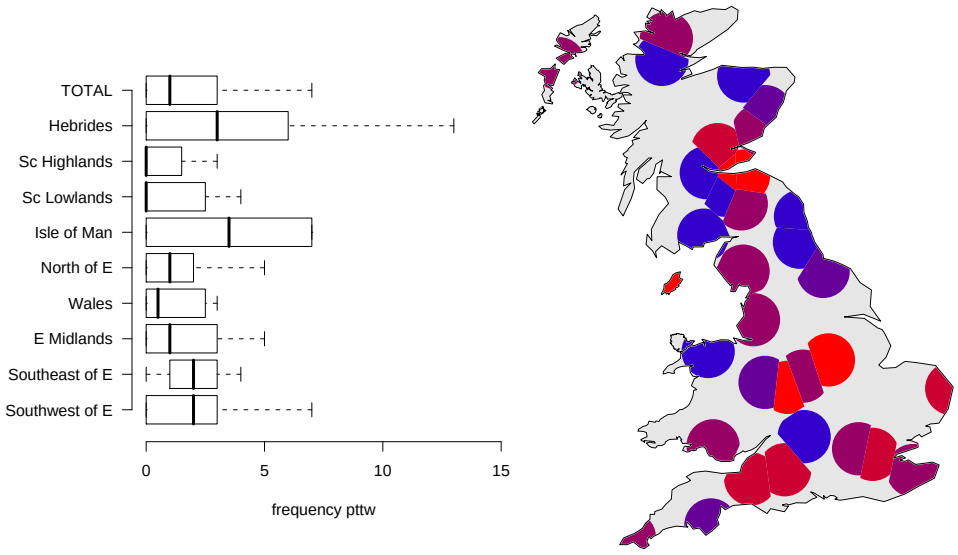

Figure 17: Feature [17] (BE GOING TO). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

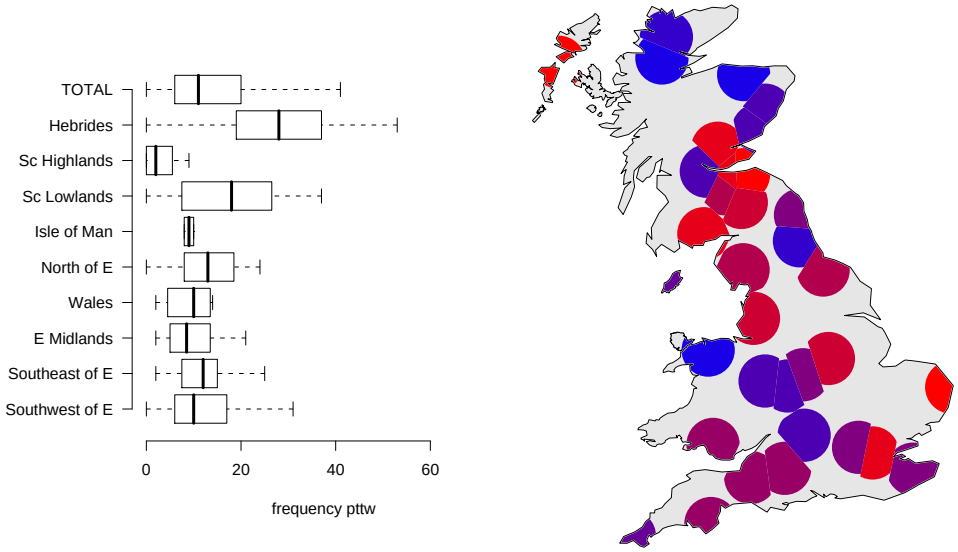

Figure 18: Feature [18] (WILL/SHALL). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.12. Habitual past: feature [19] (WOULD) / feature [20] (*used to*)

More than 2,500 instances of the forms *would* and *wouldn't*, as well as contracted forms with *'d*, were inspected as to whether they served as markers of habitual past, in which case they were manually tagged. The coding guidelines in Tagliamonte and Lawrence (2000) were followed; hence, an occurrence of WOULD had to be roughly interchangeable with *used to*, which is another way of saying that

- the occurrence had to have past time reference;
- the form had to mark "[a] situation which is characteristic of an extended period of time, so extended in fact that the situation referred to is viewed not as an incidental property of the moment but, precisely, as a characteristic feature of a whole period" (Comrie 1976: 27–28).

So, (32-a) was considered to be an instance of *would* marking habitual past, while (32-b) was not:

(32) a. <u Heb18> ... But things was – well money was very very scarce at that time, and things was cheap too. <u HebMMcl> Money *would* go farther then. <FRED HEB001>
b. Yes, I *would* say so, because the majority did get jobs in Widness ... <FRED LAN011>

A retrieval script then identified and registered

- text frequencies of habitual *would*, as in (32-a), *wouldn't* (cf. (33-a)), and *'d* (cf. (33-b)), as identified by the manual inspection procedure:

  (33) a. They had one woman at one time at Picket, you *wouldn't* be there then. <FRED YKS011)
  b. And oh, you *'d* get a party in now and again, and he would have the money. <FRED CON011>

- text frequencies of *used to* (which always functions as a habitual past marker), as in (34):

  (34) ... he *used to* go around killing pigs. <FRED NTT008>

5.13. The progressive: feature [21]

A screening script flagged all *-ing* forms in the dataset, ignoring

- all obviously non-verbal *-ing* forms (e.g. *something*, *shilling*, *nothing*, and some 14 other frequent non-verbal forms);

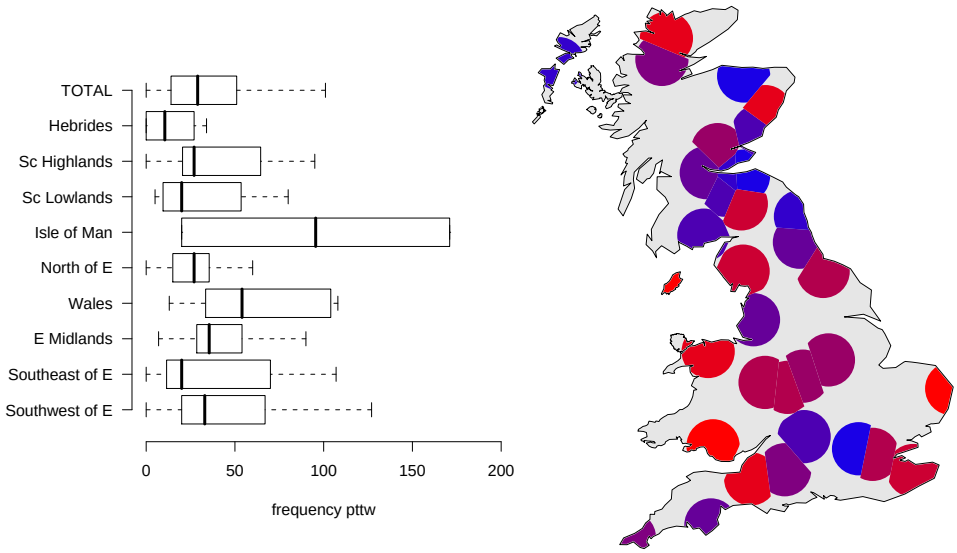

Figure 19: Feature [19] (WOULD). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

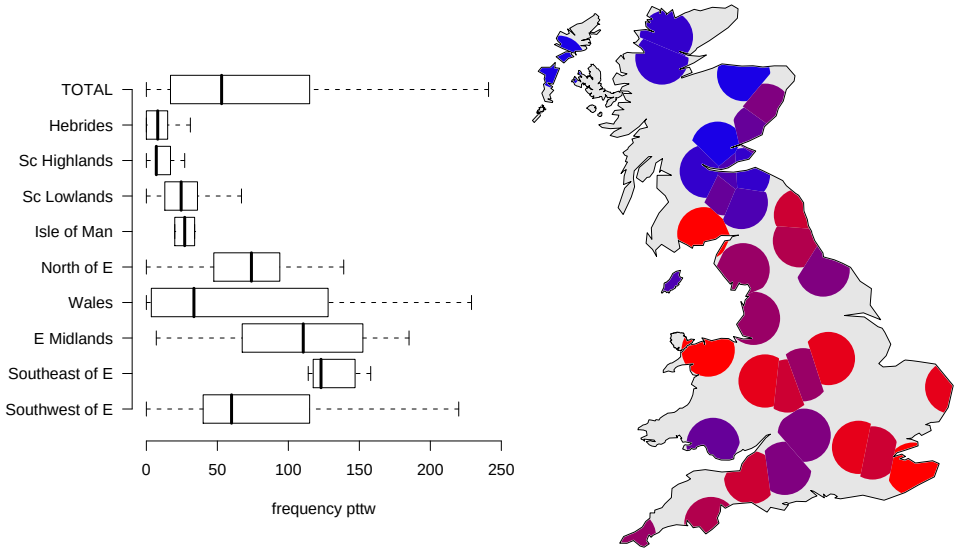

Figure 20: Feature [20] (*used to*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

- *-ing* forms that could not possibly be part of a progressive construction (e.g. tokens preceded by *the* and *a*, among some 24 other knock-out collocates).

This procedure yielded more than 4,500 tokens in the dataset where the *-ing* ending was potentially a progressive morpheme. These tokens were inspected manually/qualitatively and – following the criteria detailed in Hundt (2004: 56) – tagged if the *-ing* ending indeed coded a progressive form (be it a present progressive [as in (35-a)], a past progressive [as in (35-b)], a present perfect progressive [as in (35-c)], a past perfect progressive [as in (35-d)], or a progressive infinitive [as in (35-e)]):

(35) a. ...the rest *are going* to Portree School now ... <FRED HEB001>
b. ...and my father *was leaving* the farm he was on ... <FRED ANS001>
c. Yes, you would *have been looking* down on, on Dherue. <FRED SUT001>
d. War *had been going* on for at least two or three years ... <FRED YKS010>
e. He likes to *be joking*, Desmond. <FRED HEB001>

Tokens were not tagged if the *-ing* ending was indicative of, e.g., a gerund, as in (36-a), a form of BE GOING TO used to express futurity, as in (36-b), a construction with an adjectival participle, as in (36-c), or an appositional participle, as in (36-d):

(36) a. we did no *trawling* then ... <FRED KEN008>.
b. They wouldn't go, not to, not if they *were going to* do any damage. <FRED KEN001>.
c. That *is interesting*. <FRED ELN009>.
d. I used to go in the passing shop, *inspecting* shop with Joe Grant ... <FRED WIL007>.

Finally, a retrieval script identified and registered all *-ing* forms where the *-ing* suffix marked progressive aspect, as identified by the manual inspection procedure.

5.14. The present perfect: feature [22] (auxiliary BE) / feature [23] (auxiliary HAVE)

A screening script flagged all variant forms of the verbs TO HAVE (i.e. *have, haven't, has, hasn't, 've, 's*) and TO BE (i.e. *am, amn't, are, aren't, is, isn't, 's, 'm, 're*) in the datasets. (Note that due to comparatively high token frequencies of the HAVE perfect and comparatively low token frequencies of the BE perfect, two different datasets were analyzed: the full dataset for the BE perfect, and the abridged dataset for the HAVE perfect.) The screening script ignored the following collocational patterns:

- 43 collocational patterns where a HAVE occurrence could not possibly function as a present perfect auxiliary thanks to the absence of a past participle to the right of the occurrence: e.g., HAVE + *it / that / the / a*, as in (37):

(37) But I didn't *have it*, not miself. <FRED SOM032>

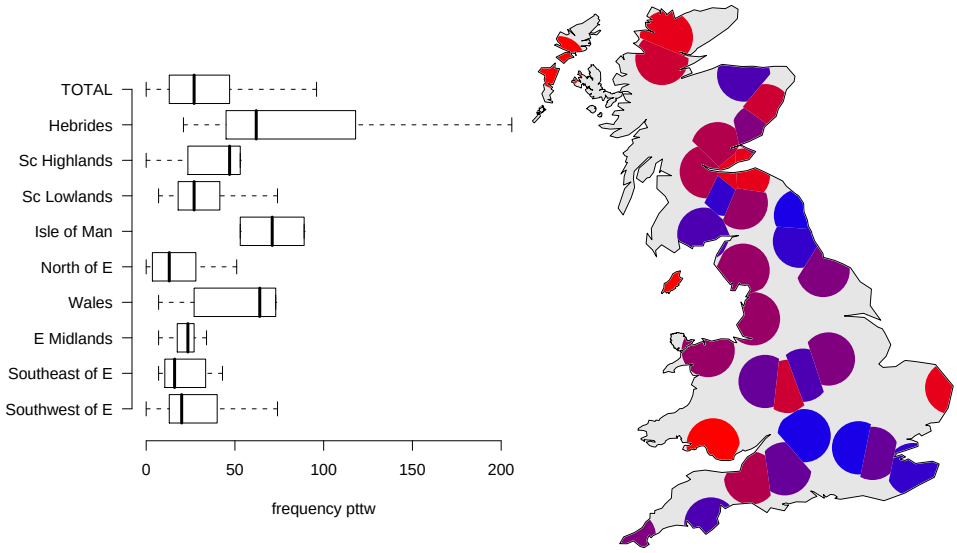

Figure 21: Feature [21] (the progressive). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

- 20 collocational patterns where a HAVE occurrence could be assumed, beyond reasonable doubt, to be a perfect auxiliary thanks to the presence of a past participle to the right: HAVE + *got* / *dealt* / *told* / *met* / *beaten* / *made* / *written* / *found* / *brought* / *built* / *sold* / *lost* / *gone* / *heard* / *known* / *thought* / *been* / *done* / *had* / *seen*, as in (38) (such occurrences were subsequently automatically retrieved):

  (38) But if you would *have seen* that job maybe you would laugh. <FRED OXF001>

- 175 collocational patterns where a BE occurrence could not function as a present perfect auxiliary thanks to, e.g., the absence of a past participle to the right of the occurrence: e.g., BE + *a* / *an* / *the*, as in (39):

  (39) Puddle Street *is the* present Crown Street. <FRED SAL006>

This procedure yielded more than 3,500 occurrences in the datasets where HAVE or BE tokens functioned as possible present perfect auxiliaries. These tokens were inspected manually/qualitatively and tagged if an immediately following past participle indicated that the HAVE or BE token indeed functioned as a perfect auxiliary, as in (40), and not, e.g., as a main verb (as indicated by the absence of a past participle), as in (41), or as a passive auxiliary, as in (42), or as a copula, as in (43):

Figure 22: Feature [22] (the present perfect with auxiliary BE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

(40) a. And that *has helped*, really. <FRED CON011>
b. ...I*'m come* down to pay the rent ... <FRED CON009>

(41) ...one of them *has* four horses and the other *has* three. <FRED WES004>

(42) ...the eight cows *are* fed from the four places, see. <FRED CON005>

(43) We don't want you, You*'re* drunk. <FRED LND005>

Whereas deciding whether a form of the verb HAVE is a perfect auxiliary is unproblematic, coding for the perfect auxiliary use of BE is not trivial. The general rule that guided the coding procedure was that if a passive (as in (42)) or adjective ((43)) interpretation was at all possible, a given BE token was *not* coded as a perfect auxiliary. Finally, a retrieval script identified and registered the relevant text frequencies.

5.15. Marking of epistemic and deontic modality: feature [24] (MUST) / feature [25] (HAVE TO) / feature [26] (GOT TO)

A retrieval script automatically identified and registered text frequencies of the following phenomena:

- MUST – the forms *must, mustn't, mustnae*, as in (44);

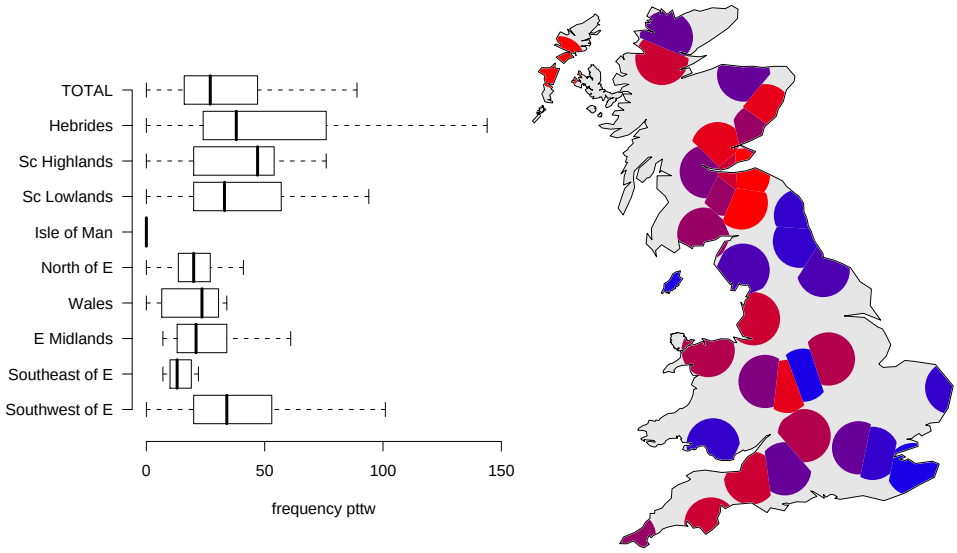

Figure 23: Feature [23] (the present perfect with auxiliary HAVE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

(44) And I *must* say that, you know, we were never home ... <FRED WIL024>

- HAVE TO – the forms *have, has, had, having, havin'* followed by the token *to*, and transcribed *hafta* and *hefta*, as in (45);

(45) Not very often did we *hafta* stop fishing. <FRED SFK020>

- GOT TO – the form *got* followed by the token *to*, and transcribed *gotta*, as in (46).

(46) I said, You *gotta* give somebody a chance somewhere. <FRED LND001>

Throughout, no distinction was made between deontic and epistemic modality.

5.16. *A*-prefixing on *-ing*-forms: feature [27]

A retrieval script automatically identified and registered all occurrences of *a*-prefixed *-ing* or *-in'* forms (as in (47)) in the dataset.

(47) a. And he was *a*-wait*ing* there, walking to and fro ... <FRED KEN003>
b. I was *a*-walk*in'* along the market one mornin' ... <FRED SFK026>

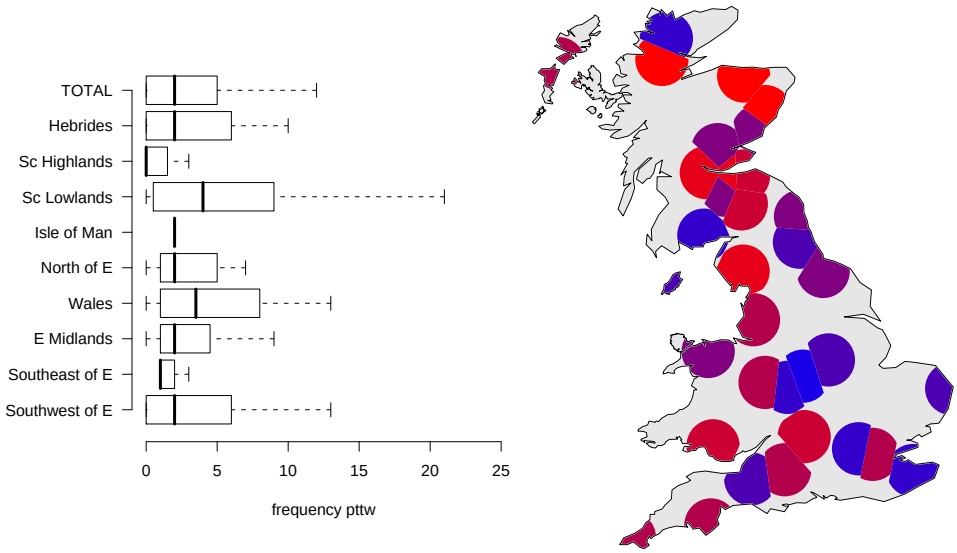

Figure 24: Feature [24] (MUST). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

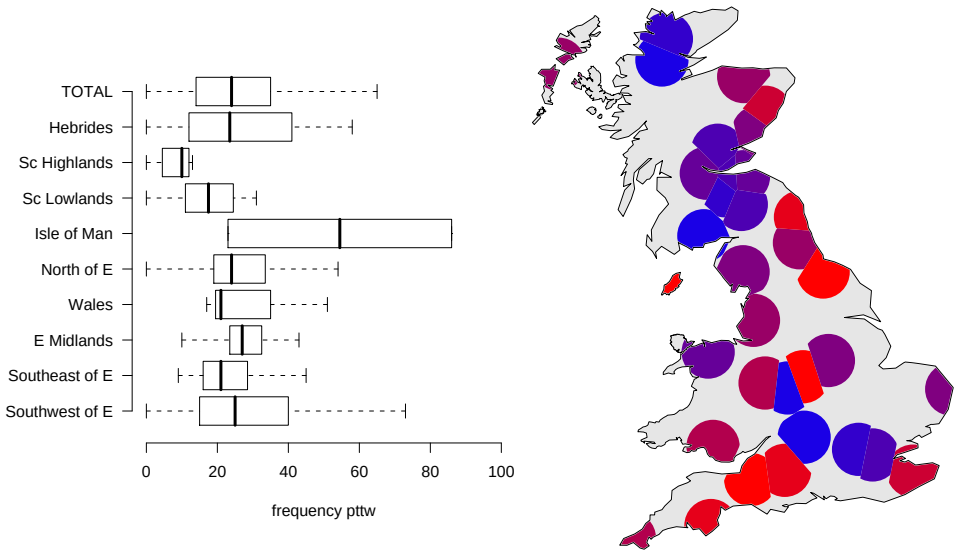

Figure 25: Feature [25] (HAVE TO). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

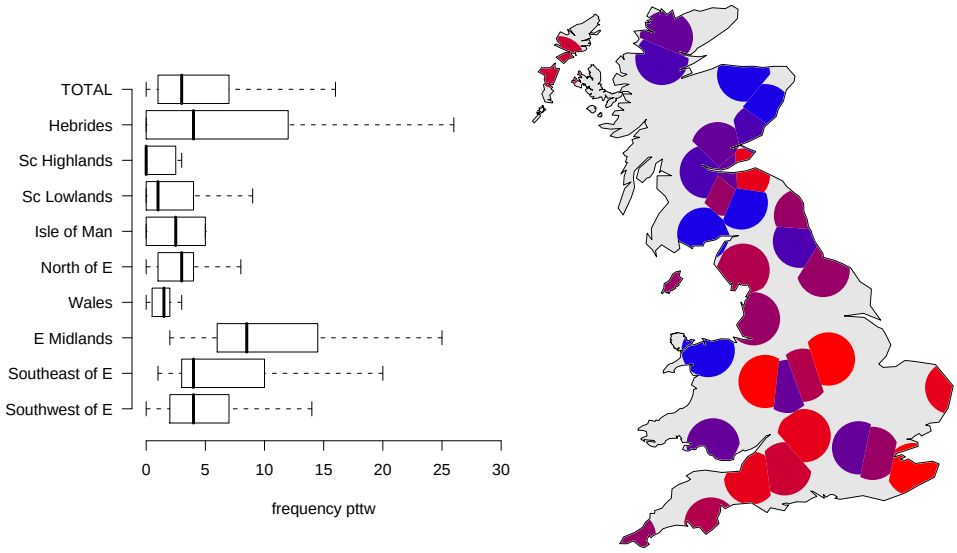

Figure 26: Feature [26] (GOT TO). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.17. Conjugation regularization: feature [28] (non-standard weak past tense and past participle forms)

A word list of *-ed* forms in the dataset was generated. In this list, all unambiguously non-standard verbal *-ed* forms – i.e., non-standard weak forms – whose frequency exceeded 10 raw hits were identified. Note that cases where the form could possibly be standard as a past tense form but not as a participle form (e.g. *show, showed, shown*) were ignored. This exercise yielded the following list of eight high-frequency non-standard weak verbs:

	non-standard *-ed* form (raw hits in FRED)	standard strong form (raw hits in FRED)
to know (*knowed* vs. *knew/known*)	108	1,483
to run (*runned* vs. *ran/run*)	35	1,476
to catch (*catched* vs. *caught*)	27	279
to draw (*drawed* vs. *drew/drawn*)	26	98
to tell (*telled* vs. *told*)	21	858
to sell (*selled* vs. *sold*)	19	549
to grow (*growed* vs. *grew/grown*)	13	186
to throw (*throwed* vs. *threw/thrown*)	14	135

Figure 27: Feature [27] (*a*-prefixing). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

Given this list, a retrieval script subsequently identified and registered aggregate text frequencies of non-standard *-ed* forms, as in (48) through (55).

(48) They *knowed* all about these things ... <FRED NTT013>

(49) Course we all *runned* away from school ... <FRED WIL003>

(50) ... so he run and he *catched* me. <FRED WES004>

(51) I never *drawed* the curtains ... <FRED SAL024>

(52) Ah, there was a fellow *telled* me that ... <FRED PEE002>

(53) Now then, we *selled* manure, four-and-six a load. <FRED YKS003>

(54) No, he just *growed* oats for his horses, see. <FRED KEN002>

(55) He *throwed* un in the fire. <FRED SOM028>

5.18. Non-standard past tense *done*: feature [29]

A screening script flagged all instances of the form *done* in the dataset, ignoring 23 collocational patterns in which *done* could not be a past tense form (as opposed to a

Figure 28: Feature [28] (non-standard weak past tense and past participle forms). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

past participle form), thanks, e.g., to a preceding auxiliary verb, as in (56):

(56) a. She used to sit there 'til everything *was done* ... <FRED WIL023>
b. ...but the driver said, What would you *have done* if we 'd been coming fast, laddo? <FRED SAL011>

This procedure yielded about 1,000 instances where the form *done* was possibly a past tense form. These tokens were inspected manually/qualitatively and tagged if *done* was indeed a non-standard past tense form. as in (57), and not, in fact, a past participle form, as in (58):

(57) a. ...you'd come home and then you'd go back again until you came home and *done* the home fishing. <FRED SFK012>
b. 'cause he *done* all his painting inside. <FRED CON004>

(58) Well, they had a lot of work *done* ... <FRED DEV005>

Finally, a retrieval script identified and registered all manually tagged occurrences of non-standard past tense *done*.

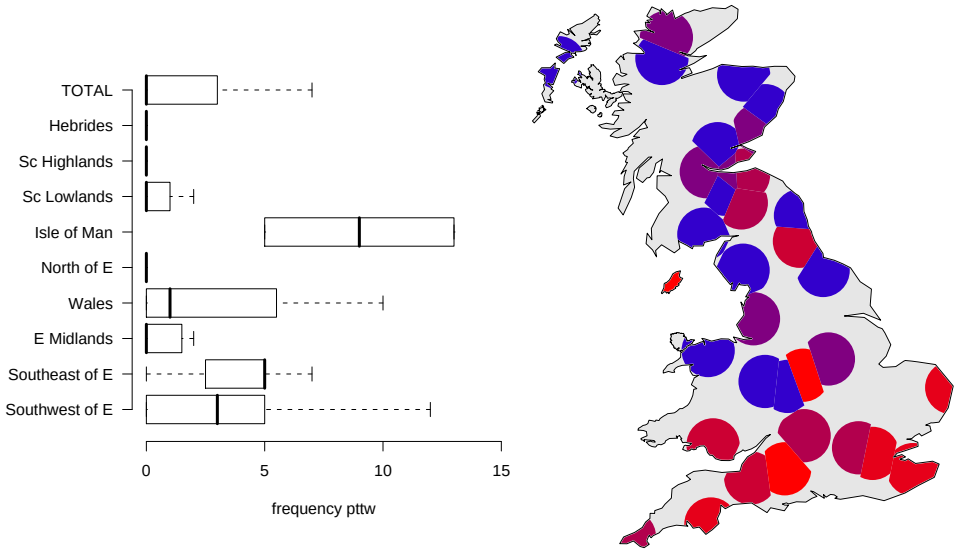

Figure 29: Feature [29] (*done*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.19. Non-standard past tense *come*: feature [30]

A screening script flagged all occurrences of the form *come* preceded by either

- the pronouns *he*, *she* or *it*, as in (59) (cf. Anderwald 2009 for a similar restriction):

(59) a. Well, out *he come*. <FRED SFK001>
b. And *she come* to have this boy ... <FRED LAN004>
c. And *it come* down over the Westbury White Horse Hill there ... <FRED WIL003>

- or by any token starting in a capital letter, as in (60):

(60) Now when old Bobby *come* in, he was going ... <FRED CON003>

This procedure yielded more than 700 tokens where the form *come* was possibly a past tense form. These tokens were inspected manually/qualitatively and tagged if *come* was indeed a non-standard past tense form, as in (61):

(61) And he *come* down the road one day, he said, Why don't you give up this place boy ... <FRED CON005>

Deciding whether any given instance of *come* is indeed a past tense form (as opposed to a present tense form or an infinitive form) can be problematic. The following guidelines were followed:

- The neighborhood of modals, as in (62), or auxiliaries/operators, as in (63), categorically ruled out coding for past tense usage:

  (62) Well, he came over here and asked this man, *could* he *come* here to work ... <FRED CON008>

  (63) ... we wondered, Well, how *did* it *come* about. <FRED CON010>

- Imperative usages, as in (64), were ignored:

  (64) ... *come* back and go back for work ... <FRED WES015>

- Whenever neighboring verb forms were clearly unmarked for past tense, as in (65), *come* forms were not coded as past tense forms either:

  (65) Well, anyhow, this submarine, when we were in action, he *come* up and *give* a burst of machine gun fire at the Telesia ... <FRED SFK003>

Subsequently, a retrieval script identified and registered the text frequencies of all manually tagged occurrences of past tense *come*.

5.20. Negative suffixes: feature [31] (*-nae*)

A retrieval script identified and registered all occurrences of tokens ending in *-nae*, as in (66):

(66) I *cannae* remember the name. <FRED NBL003>

5.21. *Ain't*: feature [32]

A retrieval script identified and registered all occurrences of the form *ain't*, as in (67):

(67) Hm, I mean people, people *ain't* got no money you see ... <FRED NTT013>

5.22. Multiple negation: feature [33]

A screening script flagged the following patterns – which constitute *possible* but not certain candidates for multiple negation – in the dataset:

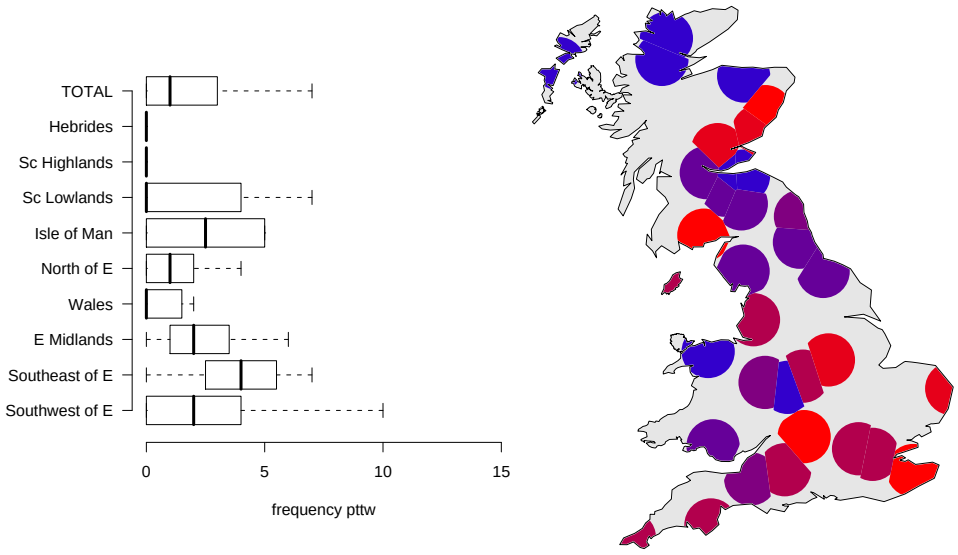

Figure 30: Feature [30] (*come*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

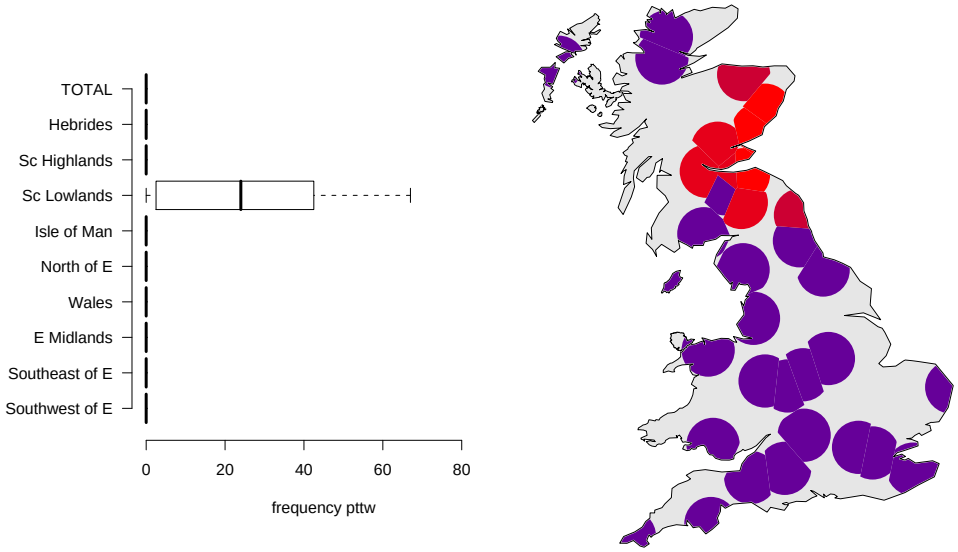

Figure 31: Feature [31] (*-nae*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

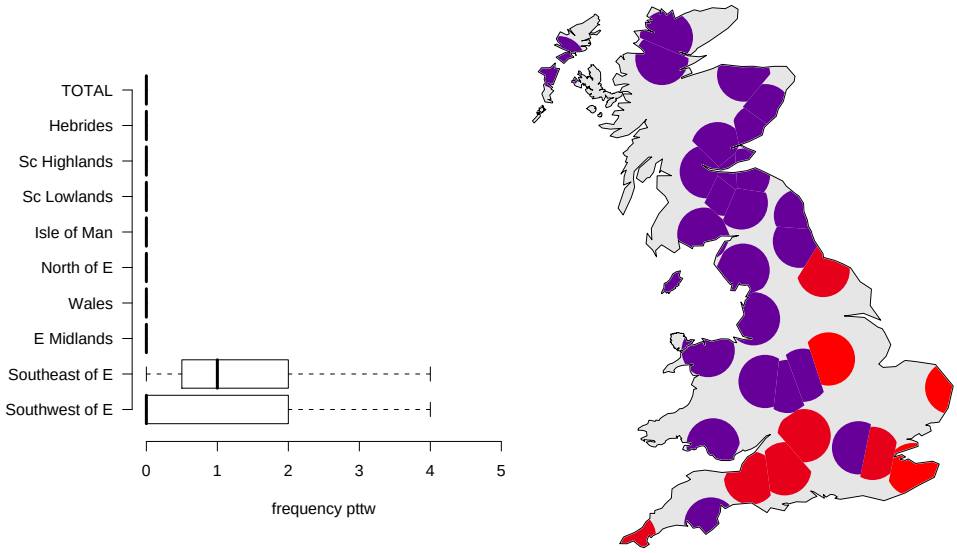

Figure 32: Feature [32] (*ain't*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

- **n't* <word> <word> *no**, as in (68) (34 instances):

(68) a. . . . do*n't* you make *no* damn mistake . . . <FRED CON005>
b. . . . it was*n't* Jude Ambry *no* it was Rigdens . . . <FRED KEN003>

- *not* <word> <word> *no**, as in (69) (37 instances):

(69) a. And she was told *not* to say *no*thing . . . <FRED DEV001>
b. . . . I mean there 's *not* so much *no*wadays. <FRED ELN014>

- *not* <word> *no**, as in (70) (57 instances):

(70) a. . . . 'cause you dare *not* say *no*thing . . . <FRED LND001>
b. It 's *not* there *no*w. <FRED LAN007>

- *never* <word> <word> *no**, as in (71) (46 instances):

(71) a. And of course he *never* come back *no* more. <dfk011>
b. You *never* see them *no*w. <FRED SFK024>

- *no** <word> **n't*, as in (72) (450 instances):

(72) a. ... *no* burning would*n't* get it out ... <FRED SOM001>
b. But *no*w they could*n't* strip them white. <FRED SOM016>

Among the above patterns in the dataset, all those occurrences were manually/qualitatively identified and tagged that indeed constitute instances of multiple negation (as in the (a) examples above), an exercise which yielded some 70 instances of genuine multiple negation. Subsequently, a retrieval script automatically retrieved these manually tagged instances, along with a number of patterns that match instances of multiple negation at all times (at least in FRED) under the condition that the expression *no** does not match the following tokens: *now, normal, nowadays, nor, noise, notes, notice, normally, Norman* (notice hat punctuation was also taken into account). Examples for such patterns include the following:

- **n't* <word> *no**, as in (73) (550 instances):

(73) Now I, I 'fraid I would*n't* know *no*thing ... <FRED CON001>

- **n't no**, as in (74) (198 instances):

(74) Well he was a, he was*n't no* fool <FRED CON003>

- *never* <word> *no**, as in (75) (293 instances):

(75) Oh, the farmer *never* said *no*thing to you about that. <FRED ANS001>

- *no* **n't*, as in (76) (53 instances):

(76) ... 'cause *no*body did*n't* have a terrible lot of cows, you know. <FRED CON005>

5.23. Contraction: feature [34] (negative contraction) / feature [35] (auxiliary contraction)

All occurrences in the dataset of tokens ending in *-n't* or *-nae*, as in (77), were automatically identified by a retrieval script, as were all instances of auxiliary contraction (*'ll not*, *'ve not*, *'s not*, *'d not*, *'m not*, and *'re not*), as in (78).

(77) a. They *won't* do anything. <FRED WES011>
b. ... but Louise is*nae* too keen on the idea ... <FRED ELN008>

(78) a. She *'ll not* know where Jerusalem Terrace is <FRED ANS004>
b. I failed biology, and I *'ve not* been to biology for the past fortnight. <FRED ELN009>

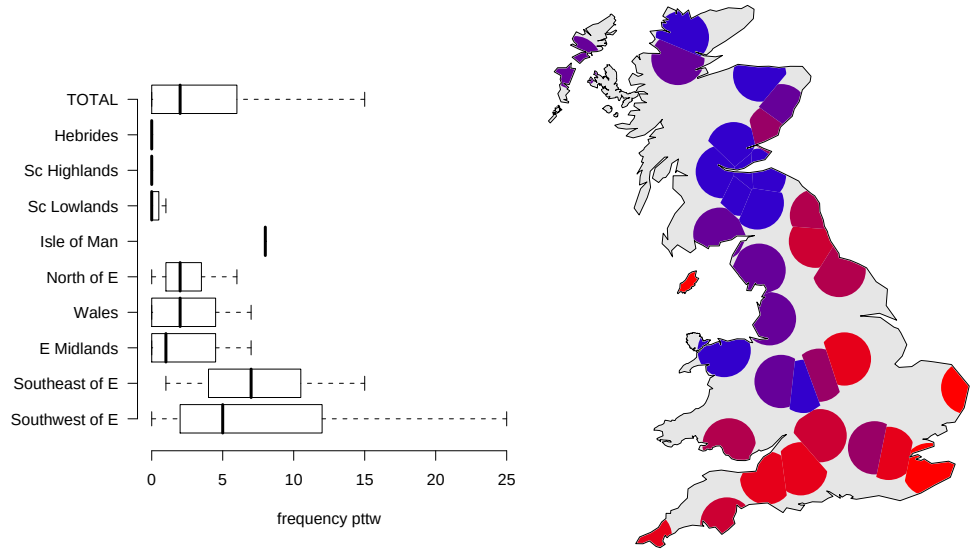

Figure 33: Feature [33] (multiple negation). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

c. That *'s not* nice. <FRED ELN009>
d. It's, it's alright, He *'d not* bother. <FRED LAN005>
e. . . . I *'m not* a racialist or anything like that, we all belong to God. <FRED LAN006>
f. We *'re not* going home tonight. <FRED LND006>

5.24. *Never* as past tense negator: feature [36]

A screening script flagged all occurrences of the form *never* in the dataset, ignoring instances where the discourse context reliably ruled out usage of *never* as a preverbal past tense negator, as in (79) (in all, there were 18 such collocational knock-out contexts):

(79) a. But I *was never* happy. <FRED HEB011>
b. I *'ve never* heard a man swear like it . . . <FRED SOM033>.
c. Oh, you *can never* tell. <FRED ANS003>

This procedure yielded more than 3,500 instances in the dataset where *never* instantiated a possible preverbal past tense negator. These occurrences were inspected manually/qualitatively and tagged if *never* indeed served as a preverbal past tense negator,

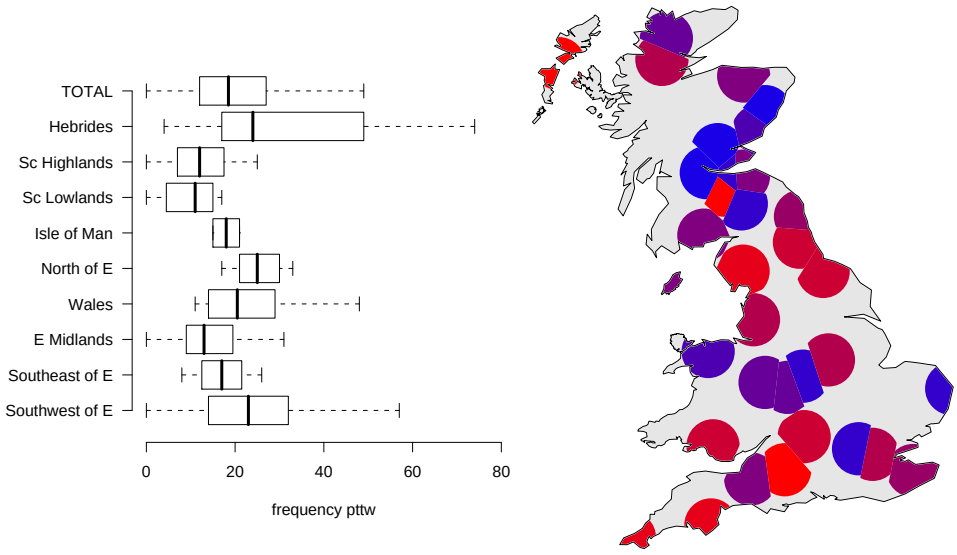

Figure 34: Feature [34] (negative contraction). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

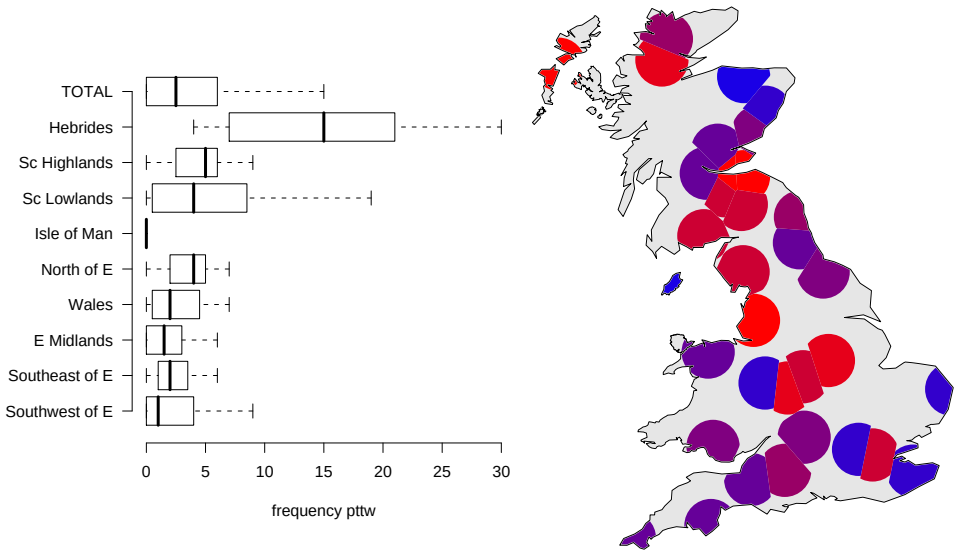

Figure 35: Feature [35] (auxiliary contraction). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

as in (80), and not simply as an adverb, as in (81):

(80) ... she'd been divorced then, she was home, see. 'cause I can't mind her married, but I can mind going down to the farm when I was a little boy, where she was still. And she was drunk. And I don't know come Uncle Den, I couldn't tell you that, you know, 'cause I wasn't more than, just like that. I was – the other put- Aunt Joan sat down there, and Kira there, and they *never moved* no more, neither one of them, *never tried* to. <FRED CON005>

(81) But no, we used to *never* give it a thought, not for that part. <FRED CON007>.

Note that it can be difficult, if not impossible, to determine whether any given occurrence of *never* indeed serves to negate a verb (which may denote a punctual event), or whether *never* has the adverbial meaning >not ever<. This is why the manual coding procedure steered clear, as far as possible, of such (necessarily subjective) semantic considerations and instead relied on formal criteria only. The coding guidelines can be summarized as follows:

1. If *never* was followed by a verb form that was explicitly marked for past tense (as in (80) above), *never* was considered a past tense negator at all times.
2. If *never* was followed by a verbal form that was not explicitly (overtly) marked for past tense, as in (82), *never* was not considered a past tense negator:

   (82) a. ... he *never put* his hand on us. <FRED LAN004>
   b. Course a horse *never does* lay down ... <FRED KEN011>

Finally, a retrieval script identified and registered text frequencies of all manually tagged occurrences of *never* as a preverbal past tense.

5.25. The *was–weren't* split: feature [37] (WASN'T) / feature [38] (WEREN'T)

All occurrences of WASN'T forms (*wasn't, wasnae*), as in (83), as well as all occurrences of WEREN'T forms (*weren't, werenae*), as in (84), were automatically identified and registered by a retrieval script.

(83) It *wasn't* very dead, no it were just busy ... <FRED LAN015>

(84) If Miss Skinner *weren't* there there was another girl used to do it. <FRED WIL023>

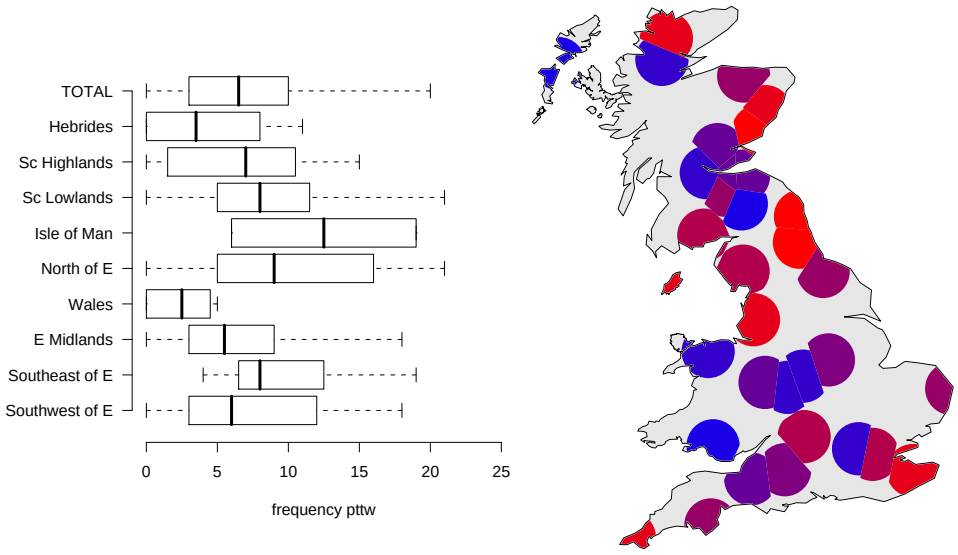

Figure 36: Feature [36] (*never* as past tense negator). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

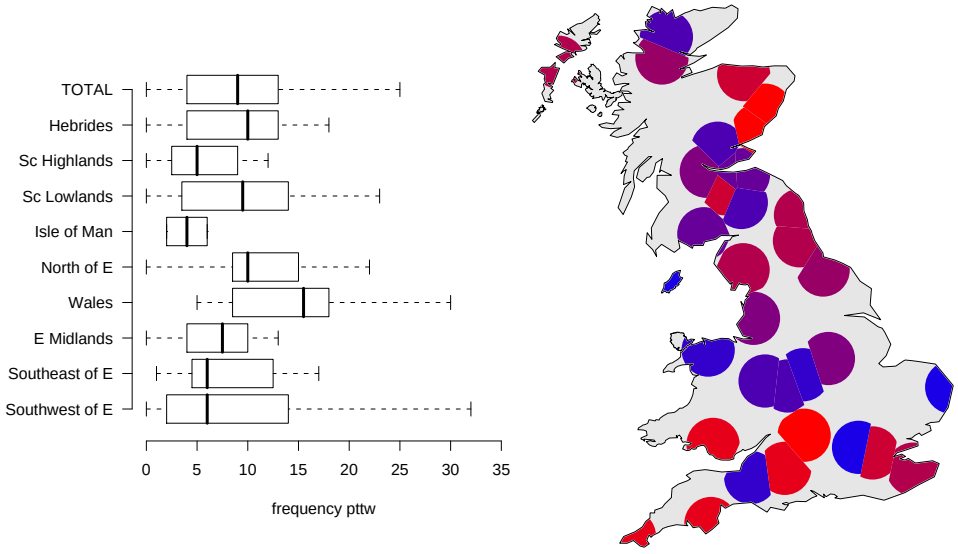

Figure 37: Feature [37] (WASN'T). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

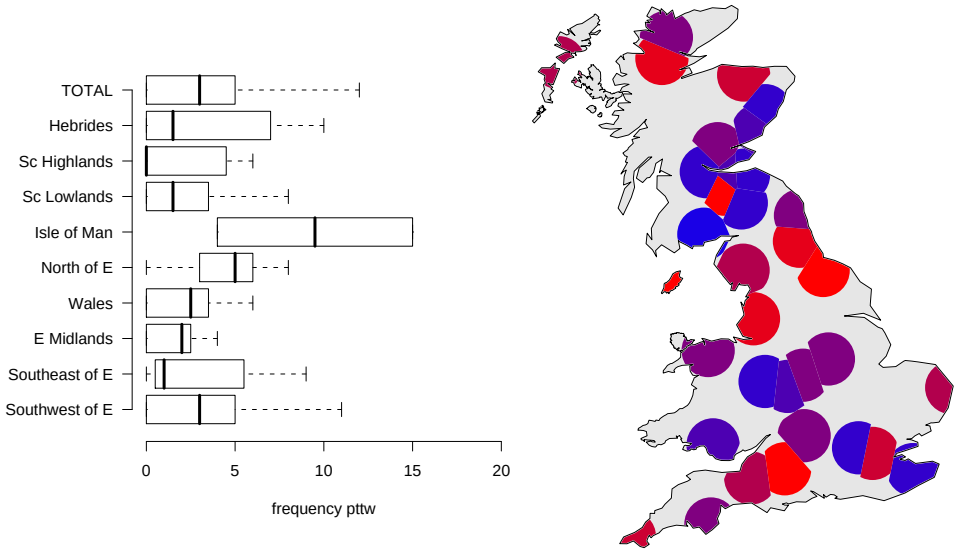

Figure 38: Feature [38] (WEREN'T). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.26. Non-standard verbal *-s*: feature [39]

A word list of all tokens in FRED ending in the grapheme <s> was generated. In this list, all clearly (cf. (85-a)) or possibly non-verbal forms (cf. (85-b) vs. (85-c)) were manually identified and saved in an exclusion list.

(85) a. There were all these *youngsters*. <FRED WES006>
b. ...he goes round doing clocks and *watches*. <FRED SAL035>
c. ...he *watches* too much television. <FRED HEB039>

Subsequently, a retrieval script identified and registered text frequencies of all tokens (i) not in the exclusion list, (ii) ending in the grapheme <*s*>, and (iii) preceded by non-3rd-person-singular pronouns (i.e. *I*, *you*, *we*, or *they*), as in (86).

(86) So *I says*, What have you to do? <FRED LAN001>

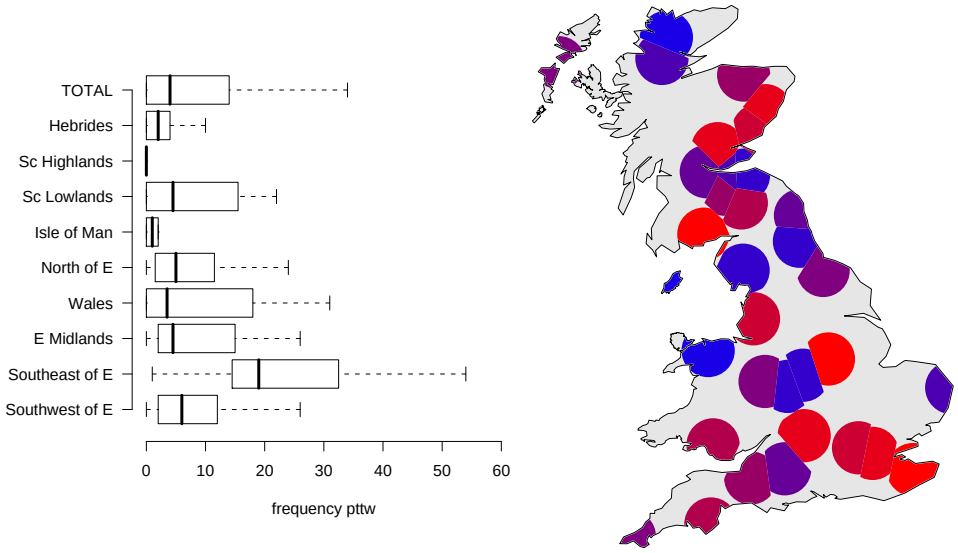

Figure 39: Feature [39] (non-standard verbal *-s*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.27. *Don't–doesn't* variability: feature [40] (DON'T with 3rd person singular subjects) / feature [41] (DOESN'T)

A screening script flagged all occurrences of the form *don't* in the dataset, ignoring (obviously irrelevant) cases where *don't* was preceded by the pronouns *I, you, we*, and *they*, as in (87):

(87) a. *I don't* know what to say ... <FRED NTT001>
b. Games as *you don't* see today, piggy, jump on back, jumping on back ... <FRED LAN020>
c. Ach *we don*'t get much snow here anyway, not really. <FRED HEB038>
d. 'Cause *they don't* use that cemetery now you see, it's finished. <FRED DEN001>

This exercise yielded some 800 instances of *don't* in the dataset where the subject was possibly in the 3rd person singular. These tokens were inspected manually/qualitatively and tagged if the subject was indeed in the 3rd person singular, as in (88-a), and not in any other person or number, as in (88-b):

(88) a. ...and then I'll send thee a letter, if *this man don't* come up to it <FRED SOM032>

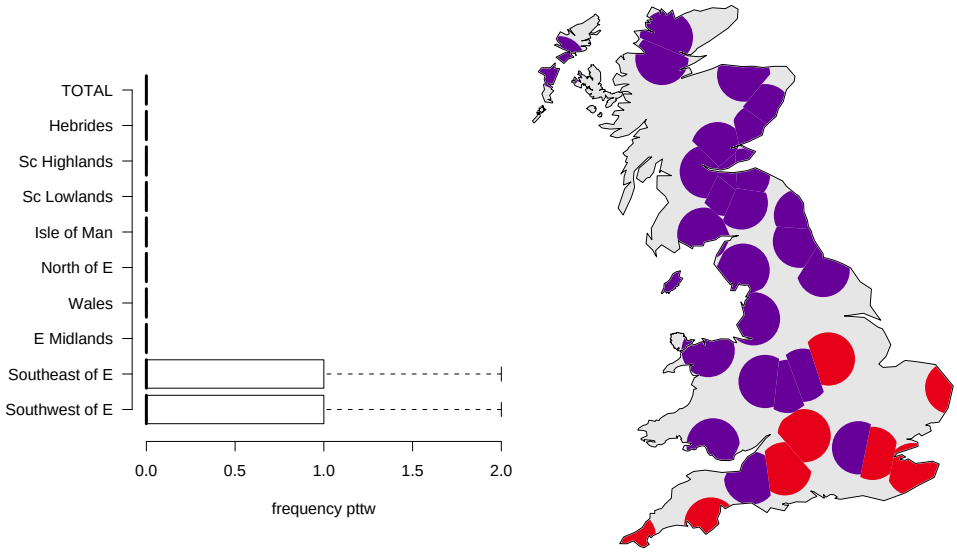

Figure 40: Feature [40] (DON'T with 3rd person singular subjects). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

b. ... *don't you* make no damn mistake ... <FRED CON005>

Finally, a retrieval script identified and registered text frequencies of

- all manually tagged occurrences of *don't* with 3rd person singular subjects;
- all DOESN'T occurrences (i.e. *doesn't*, *doesnae*), as in (89).

(89) a. Charcoal that 's right, he *doesn't* know how to burn a barrel ... <FRED SOM001>
b. He *doesnae* take an interest in anybody else except they two. <FRED ELN012>

5.28. Existential/presentational *there is/was* with plural subjects: feature [42]

A screening script flagged all occurrences of the sequences *there's, there is, there was, there isn't, there wasn't, there isnae*, and *there wasnae* in the dataset, ignoring cases where the immediately following material was obviously a singular NP, as in the examples in (90):

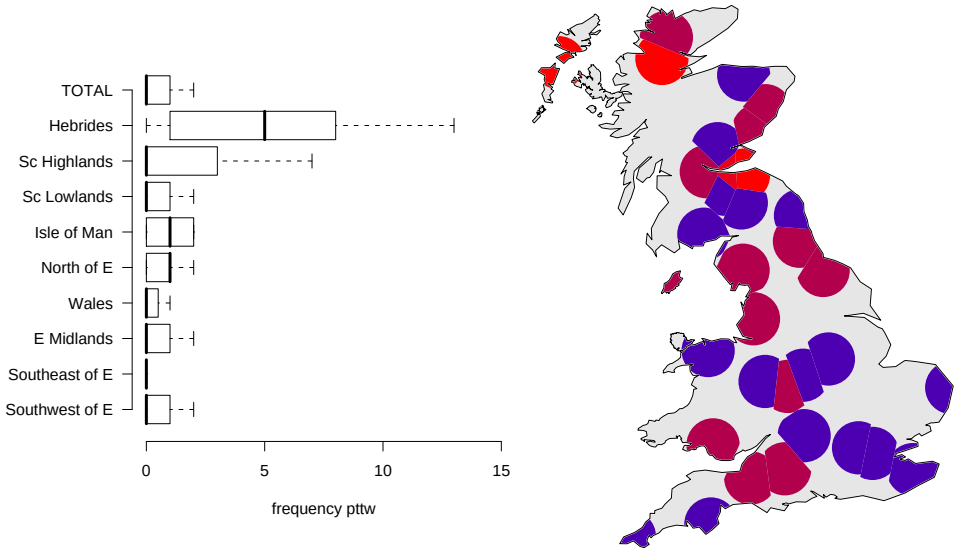

Figure 41: Feature [41] (DOESN'T). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

(90) a. I believe *there was a* Gospel Hall down there. <FRED SAL024>
b. In fact *there was one* Canadian soldier came ashore in full kit ... <FRED SUT002>.

This procedure yielded over 5,000 *there* + BE sequences in the dataset possibly followed by a plural NP. These sequences were inspected manually/qualitatively and tagged if a plural NP indeed followed, as in (91-a), as opposed to a singular NP, as in (91-b). Observe that if there were two or more conjoined singular NPs, as in (91-c), no tag was inserted.

(91) a. ... *there was thirty-seven children* involved. <FRED LAN018>
b. But eh, oh no, *there wasn't such a thing* as paper. <FRED ANS001>.
c. ... *there was me*, *him*, and *my father* ... <FRED DUR001>

Finally, a retrieval script identified and registered all manually tagged occurrences of *there* + BE followed by a plural NP.

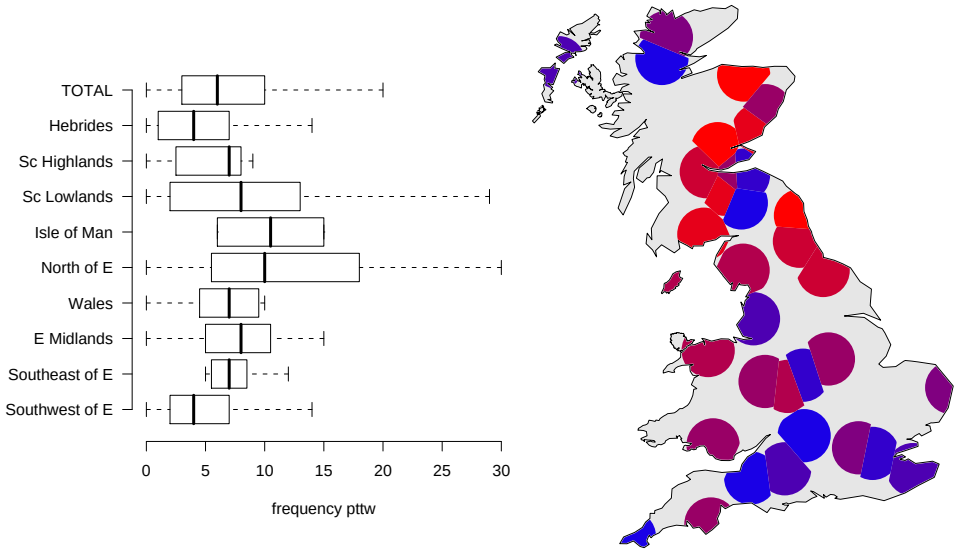

Figure 42: Feature [42] (existential/presentational *there is/was* with plural subjects). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.29. Copula/auxiliary absence: feature [43] (absence of auxiliary BE in progressive constructions)

A screening script flagged all occurrences in the dataset of the pronouns *I, you, he, she, it, we, they* that were followed by a token ending in *-ing* or *-in'*, unless the pronoun itself was preceded by contracted forms such as *'s*, *'re*, or *n't*, as in (92):

(92) ... and he said, are*n't you* go*ing* to go to work? <FRED IOM001>

This procedure yielded some 600 pronoun + **ing/*in'* sequences in the dataset that instantiated possible instances of progressive constructions where auxiliary BE was omitted. These occurrences were inspected manually/qualitatively and tagged if they were, indeed, relevant, as in (93):

(93) I said, How *you doing*? <FRED CON006>

Occurrences were not tagged if the pronoun + **ing/*in'* sequence was, e.g., a non-finite complement clause, as in (94):

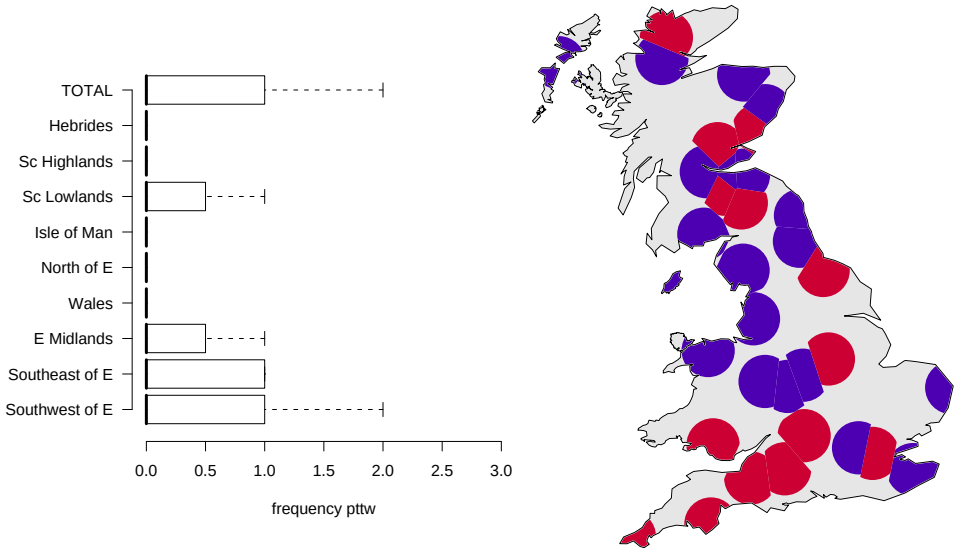

Figure 43: Feature [43] (absence of auxiliary BE in progressive constructions). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

(94) ...and then after a while I saw [*it* crawling across the gateway]. <FRED CON002>

Finally, a retrieval script identified and registered all manually tagged occurrences of progressive constructions without an auxiliary.

5.30. *Was/were* variation: feature [44] (non-standard WAS)

A screening script flagged all instances of WAS (i.e. *was* and *wasn't* or *wasnae*) in the dataset, ignoring cases where the subject was obviously a 1st or 3rd person singular subject, as indicated by a corresponding, immediately preceding pronoun or dummy subject (*I, he, she, it, that, there*; cf. (95)):

(95) My brother come home. *He was* in the army, he come home for a weekend. <FRED KEN002>

This exercise yielded about 3,000 occurrences of WAS in the dataset where the subject was possibly *not* in the 1st or 3rd person singular. These tokens were inspected manually/qualitatively and tagged if the subject was indeed not in the 1st or 3rd person singular, as in (96-a), as opposed to cases where it was (as in (96-b)). Note that if the subject was a collective noun, as in (97), no tag was inserted.

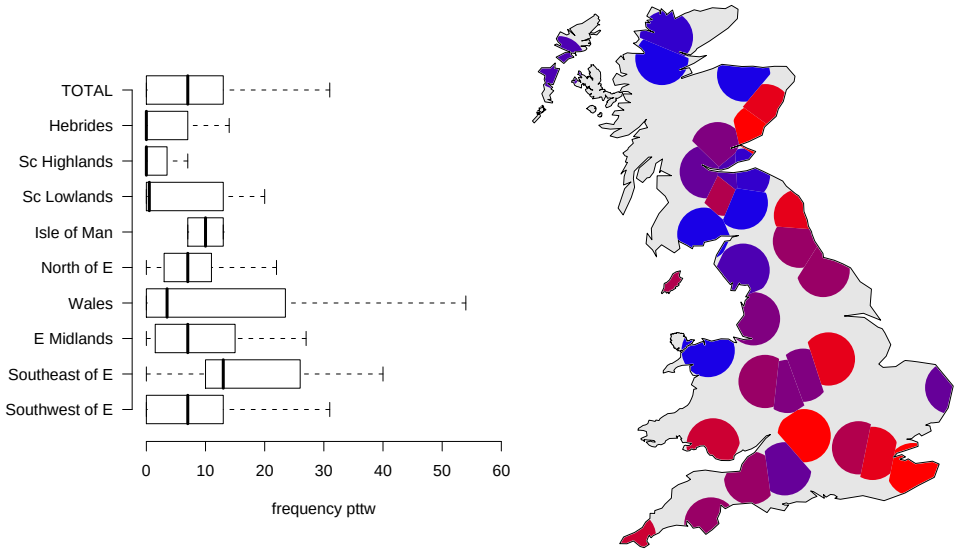

Figure 44: Feature [44] (non-standard WAS). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

(96) a. ...three of them came back. Three of them *was* killed. <FRED WLN004>
b. But my father *wasnae* originally a fisherman. <FRED BAN001>.

(97) *The police was* bitter against black-legging. <FRED GLA003>

Finally, a retrieval script identified and registered all manually tagged occurrences of non-standard WAS.

5.31. *Was/were* variation: feature [45] (non-standard WERE)

A screening script identified all instances of WERE (i.e. *were* and *weren't* or *werenae*) in the dataset, ignoring cases where WERE obviously had a 2nd person singular or a plural subject, as indicated by a corresponding, immediately preceding pronoun (*you, we, they, these, those*; cf. (98)).

(98) ...he had brothers as well, *they were* all in the trawler business you see ... <FRED YKS007>

This exercise yielded some 900 instances of WERE in the dataset where the subject was possibly in the 1st or 3rd person singular. These tokens were inspected manually/qualitatively and tagged if the subject was indeed in the 1st or 3rd person singular,

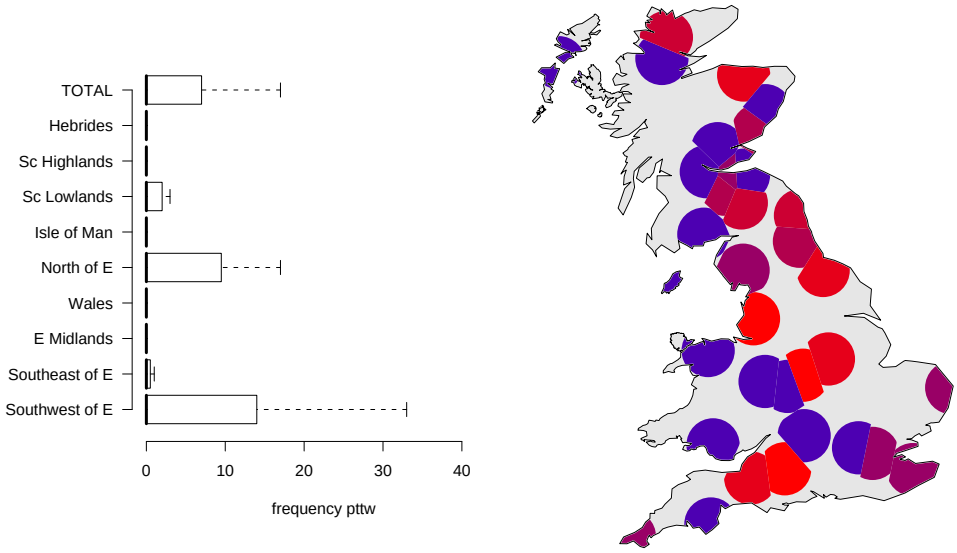

Figure 45: Feature [45] (non-standard WERE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

as in (99-a), as opposed to cases where it was not, as in (99-b). Note that if the subject was a collective noun, as in (100), no tag was inserted.

(99) a. He run away from home when he *were* a young lad. <FRED LAN005>
b. Jobs *were* scarce and he got one up the country. <FRED LAN011>.

(100) Oh yes and *the police were* watching you in the morning going to work and all, yes. <FRED GLA002>

Finally, a retrieval script identified and registered all manually tagged occurrences of non-standard WERE.

5.32. Relativizer choice: feature [46] (*wh*-relativization) / feature [47] (relative particle *what*) / feature [48] (relative particle *that*)

A screening script flagged all occurrences of the potential relativizer tokens *what*, *that*, and *who/which/whose* in the dataset (note that the *wh*-form *whom*, "which appears rarely if ever in regional dialects" [Sanderson and Widdowson 1985: 35], was not considered), ignoring in all 27 frequent collocational patterns that reliably rule out the relativizer usage of these tokens – for instance, the sequence *wondered what*, as in (101-a), or the sequence *when that*, as in (101-b):

(101) a. . . . they all *wondered what* it was. <FRED SEL001>
b. No, can't remember *when that* come. <FRED NTT009>

This procedure yielded more than 7,000 instances in the dataset where the above tokens possibly introduced a relative clause. Subsequently, the tokens were inspected manually/qualitatively and tagged if the above tokens indeed introduced a relative clause – which is another way of saying that the token had to introduce a dependent clause that postmodifies an NP and that contains a finite VP, as in (102). No distinction was made between restrictive relative clauses, as in (103-a), and non-restrictive relative clauses, as in (103-b). Needless to say, other functions of the target tokens – for instance, demonstrative *that*, as in (104) – were not coded.

(102) a. He never had any of the money *what* he earnt. <FRED KEN010>
b. The highest number *that* I can remember, I think, was fifty-two . . . <FRED CON007>
c. . . . and it was the poor people *who* poached usually. <FRED SOM019>
d. . . . mother's killed a rabbit for Sunday dinner *which* weighed twenty-two pounds. <FRED SAL006>
e. Now we had a chap *whose* name was Brown . . . <FRED SAL028>

(103) a. . . . and there was a chap *who* looked after the boilers . . . <FRED WES012>
b. My father sold it to Gerry Falkner who now owns a lot of race horses . . . <FRED SOM033>

(104) a. They were the principal blacksmiths in Forfar *at that* time. <FRED ANS003>

Finally, a retrieval script identified and registered all manually tagged relativizer occurrences.

5.33. *As what* or *than what* in comparative clauses: feature [49]

A retrieval script identified and registered all occurrences in the dataset of the sequences *as what* (cf. (105-a)) and *than what* (cf. (105-b)).

(105) a. And it, I saw it as many times *as what* the operators did. <FRED YKS001>
b. . . . well we we done no more *than what* other kids used to do . . . <FRED LEI002>

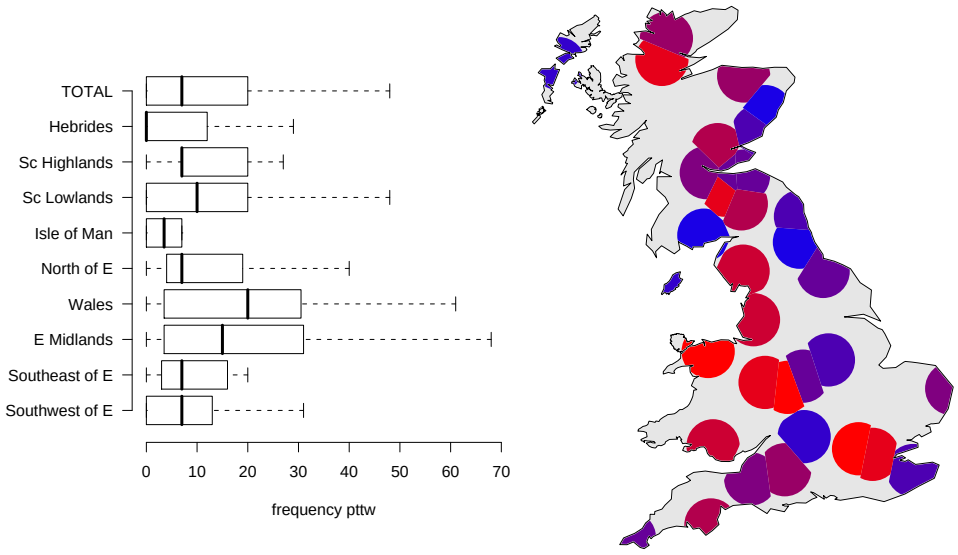

Figure 46: Feature [46] (*wh*-relativization). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

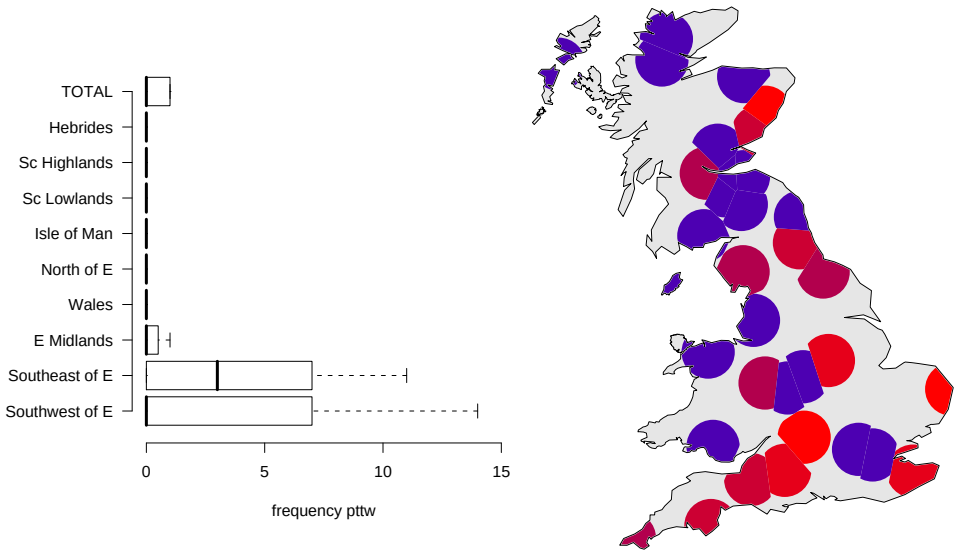

Figure 47: Feature [47] (relative particle *what*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

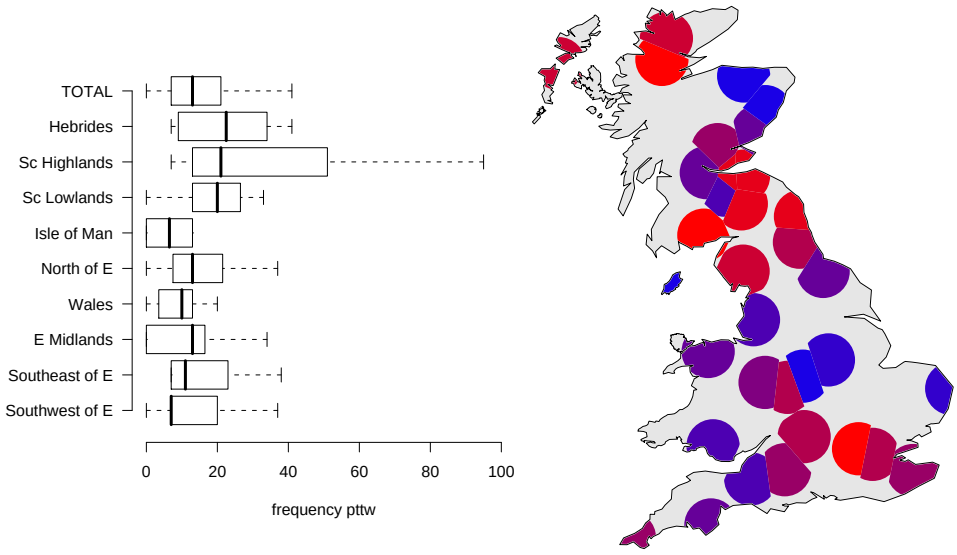

Figure 48: Feature [48] (relative particle *that*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

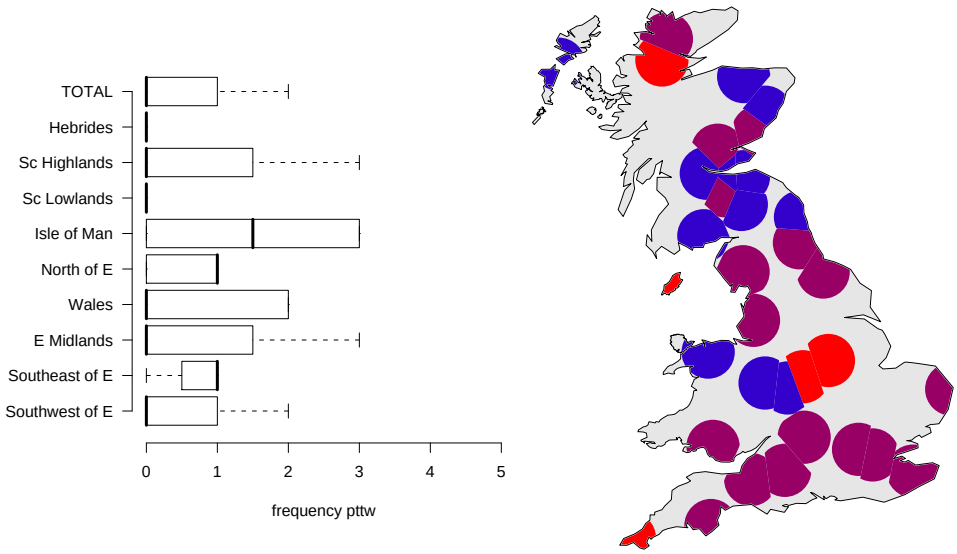

Figure 49: Feature [49] (*as what* or *than what* in comparative clauses). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

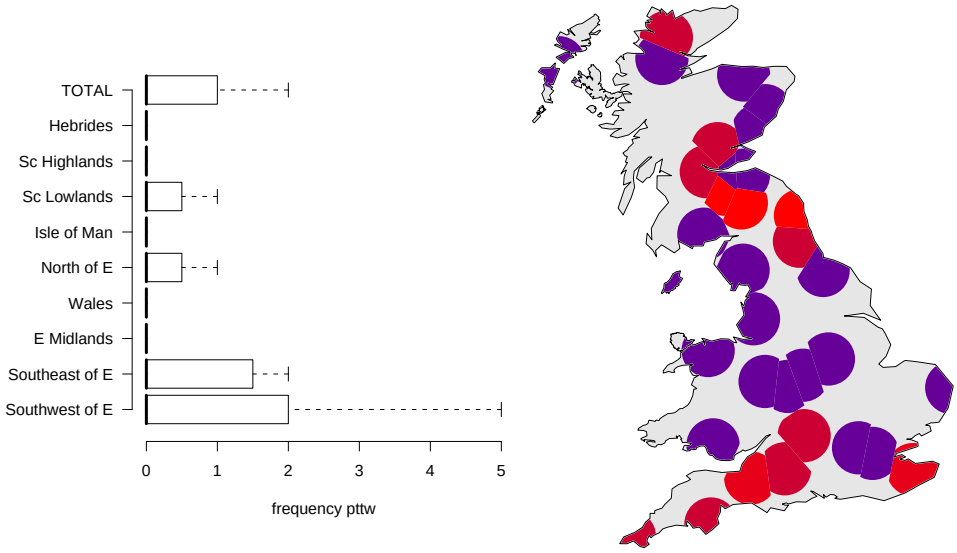

Figure 50: Feature [50] (unsplit *for to*). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.34. Unsplit *for to*: feature [50]

All occurrences of the sequence *for to* in the dataset were identified and registered by a retrieval script. No distinction was made between complement clauses (as in (106-a)), postmodifying clauses (as in (106-b)), or adverbial clauses (as in (106-c)) introduced by *for to*.

(106) a. It was ready *for to* go away with the order, or whatever. <FRED SAL016>
b. . . . and one *for to* look after the bundles as them come out the other end. <FRED CON007>
c. And *for to* launch him they used to take him over and go down the slipway. <FRED SOM036>

5.35. Verbal complementation after BEGIN, START, CONTINUE, HATE, and LOVE: feature [51] (infinitival complementation) / feature [52] (gerundial complementation)

Following the methodology utilized in Szmrecsanyi (2006: 155–156), the following tokens were identified in the dataset: forms of TO BEGIN (*begin, begins, beginning, began, begun*), START (*start, starts, starting, started*), CONTINUE (*continue, continues, con-*

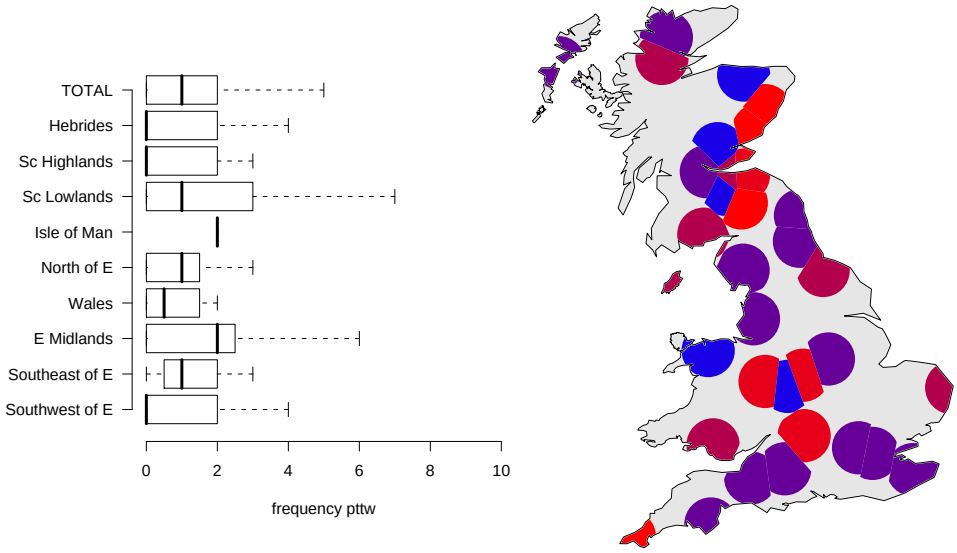

Figure 51: Feature [51] (infinitival complementation). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

tinuing, continued), HATE (*hate, hates, hating, hated*), and LOVE (*love, loves, loving, loved*). Subsequently, a retrieval script automatically identified and registered all those occurrences of the above verb forms which were followed

- by either the token *to* (infinitival complementation), as in (107);

  (107) And then I *began to* take an interest in it ... <FRED DFS001>

- or by a token ending in *-ing* (gerundial complementation), as in (108).

  (108) ... as things were getting worse they *began* call*ing* up the younger agricultural workers ... <FRED CON008>

5.36. Clausal complementation after THINK, SAY, and KNOW: feature [53] (zero complementation) / feature [54] (*that* complementation)

A screening script flagged all occurrences of the verbs THINK (*think, thinks, thinking, thought*), SAY (*say, says, saying, said*), and KNOW (*know, ken, kenned, kent, kens, knows, knowing, knew, knowed*) in the dataset, ignoring certain frequent collocations not relevant to the alternation at hand – for instance, if the verbal form was followed

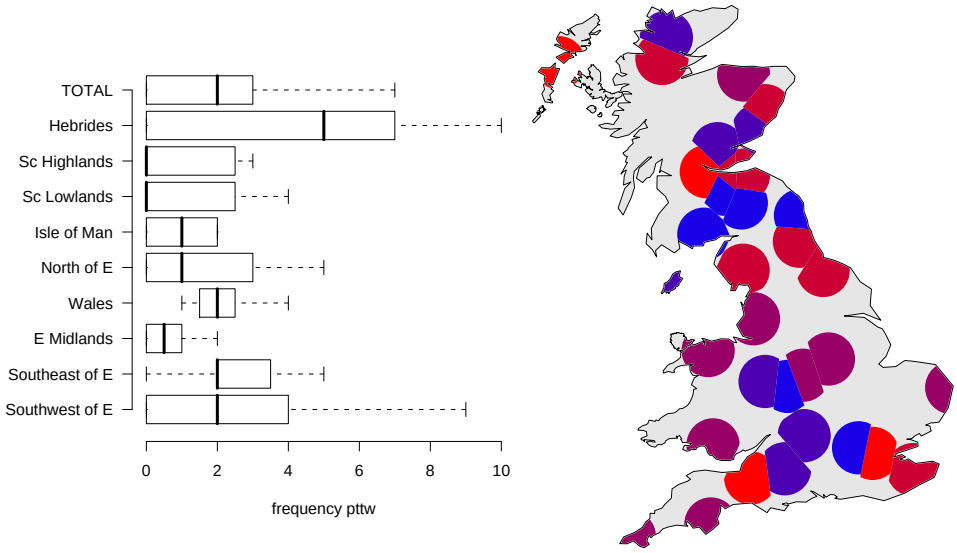

Figure 52: Feature [52] (gerundial complementation). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

by *about*, as in (109), or *of*, as in (110), the token was ignored (in all, there were 31 such knock-out collocations):

(109) I don't *know about* the money ... <FRED HEB023>

(110) I tried to *think of* it the other night. <FRED LND001>

This procedure yielded more than 10,000 potential candidates for *that* or zero complementation in the dataset. These candidates were inspected manually/qualitatively and tagged if a complement clause – either introduced by the complementizer *that*, as in (111), or by no overt complementizer, as in (112) – followed the verbal form:

(111) a. I mean they just *thought* [*that* it isn't for girls]. <FRED SAL029>
b. No, they just *said* [*that* you were there]. <FRED HEB012>
c. But, I don't *know* [*that* anybody had television]. <FRED LAN011>

(112) a. No I don't *think* [they ever thought of it]. <FRED YKS008>
b. But he *said* [it was never the same] ... <FRED KEN004>
c. And they *knew* [they'd get away with it]. <FRED LND007>

The manual coding procedure was guided by the following guidelines:

1. *'In dubio pro zero':* zero was 'default', evidence was required for *that* complemen-

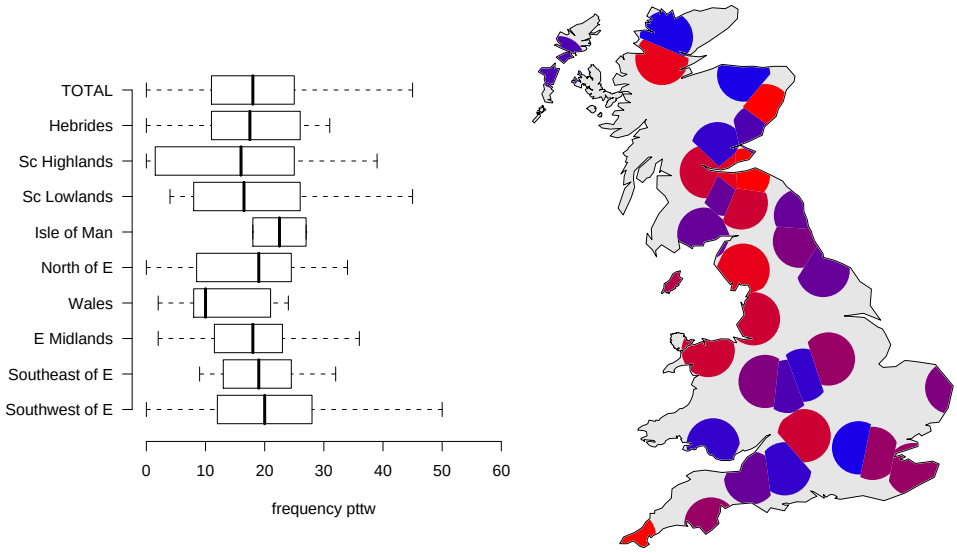

Figure 53: Feature [53] (zero complementation). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

tation.

2. *The notion of 'clause':* a complement clause was considered present only if a subject and a verb (though not necessarily a complete predicate) were present.
3. *False starts, corrections, etc.:* what counted was the final version.
4. *Punctuation is never decisive:* if everything except punctuation seemed to indicate the existence of a clause/no clause, punctuation was ignored.
5. *Direct speech* (SAY)*/direct thought* (THINK)*:* the existence of a complement clause was ruled out by the presence of at least two of the following four features: (i) lack of shifts (tense back shifts / shifts of deictic expressions / pronoun shift), (ii) interjections (*Oh*, *Well*, swear words), (iii) a comma after the matrix verb, and (iv) an upper case first letter. In the case of THINK, the collocations *think to myself* or *used to think* also ruled out coding for a complement clause.

Finally, the manually inserted tags were automatically identified and registered utilizing a retrieval script.

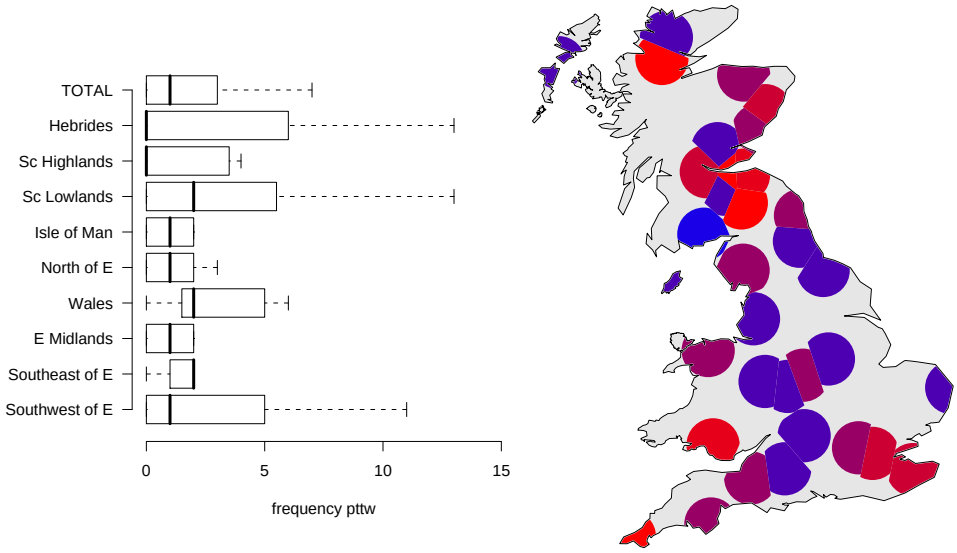

Figure 54: Feature [54] (*that* complementation). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

5.37. Lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes/no*-questions: feature [55]

A screening script identified all transcribed question mark characters (<?>) in the data set, ignoring 15 frequent patterns that categorically rule out the presence of a genuine *wh*-question or main clause *yes/no*-question – for instance, the sequence **n't it?* (indicative of a tag question), as in (113), or set phrases such as *shall I say?*, as in (114):

(113) It's hard work *isn't it?* <FRED HEB014>

(114) But the person, the mate would turn round and then he would have more or less, what *shall I say?* <FRED SFK020>

This procedure yielded some 4,000 instances where the presence of a transcribed question mark possibly instantiated the target phenomenon. Subsequently, these occurrences were inspected manually/qualitatively and tagged if the question mark character indeed followed either a *wh*-question (as in (115)) or a main clause *yes/no*-question (as in (116)), *and* if the clause lacked inversion (as in (115-a) and (116-a)) and/or an auxiliary (as in (115-b) and (116-b)).

(115) a. She says, *how long you've been out?* <FRED NTT002>

b. ... but, *where you put the shovel?* <FRED CON005>

(116) a. He says I'm going to let you home, there's no sign of the stone anyway, he says. *And you're feeling allright?* Yes says I, I'm feeling fine. <FRED HEB018>
b. ... *you know that song?* <FRED DEV008>

As for the guidelines for the coding procedure, note that the classification was based on purely formal, and not semantic, criteria. For instance, (116-a) arguably carries some suggestive force which was nonetheless ignored. Note also that verbless (as in (117-a)) and/or subjectless (as in (117-b)) question clauses were ignored, as were set phrases (as in (118)) and tag questions that were not filtered out at the outset (as in (119)).

(117) a. *What, the second war?* <FRED DEV007>
b. ... then you would get the number of sections to go on your bar. *Got it?* <FRED WIL022>

(118) I beg your pardon? <FRED YKS003>

(119) I says, You asked me to do it, *didn't you?* <FRED NTT001>

A retrieval script subsequently identified and registered all tagged target phenomena.

5.38. The dative alternation after GIVE: feature [56] (prepositional datives) / feature [57] (double object constructions)

A screening script flagged all variant forms of the verb GIVE (*give, gives, gave, given, giving, givin', gived*) in the dataset, ignoring obviously monotransitive usages as indicated by the presence of the particles *up*, *away*, or *out*, as in (120).

(120) a. And I had to *give up* smoking, because I couldn't afford to buy them! <FRED LAN008>
b. ... we used to *give* it *away* for scones. <FRED WES007>
c. ... and the Manager started *giving out* free drinks ... <FRED DEV006>

This procedure yielded more than 2,000 instances of the verb GIVE in the dataset, all of which were inspected manually/qualitatively and tagged if they were part of a prepositional dative construction, as in (121-a), or a double object construction, as in (121-b).

(121) a. And, uh, he had a scheme there, the headmaster, where he used *to give three penny piece to the class* that was hundred percent all the week. <FRED LND004>

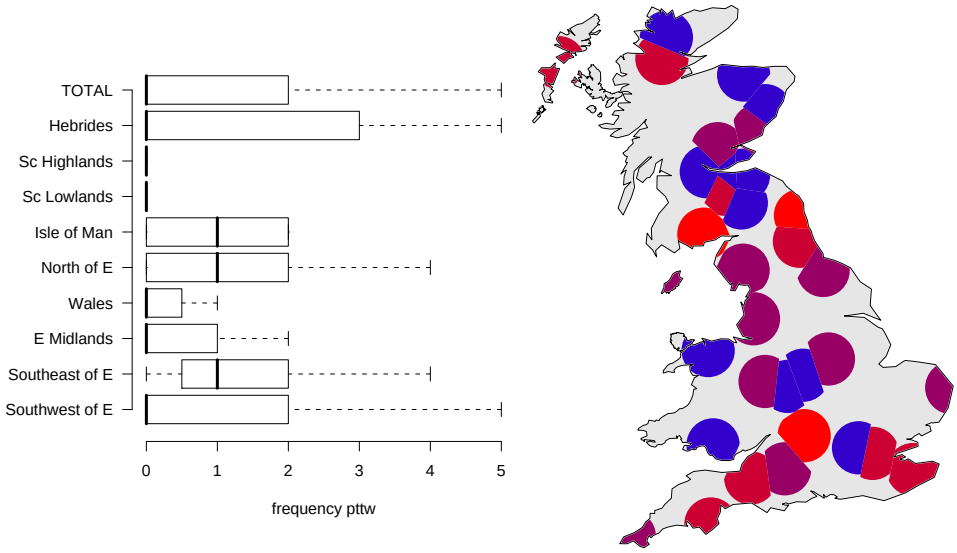

Figure 55: Feature [55] (lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes*/*no*-questions). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

b. . . . and eh I got a job and then she *gave me a job* in the warehouse. <FRED WIL007>

The manual annotation process drew on the following, more specific guidelines:

1. Passive forms of the verb GIVE were coded if their active counterparts yielded a prepositional dative structure, as in (122-a), or a double object structure, as in (122-b):

   (122) a. . . . so she gave it up and *it was given to us* . . . (∼ they gave it to us) <FRED HEB001>
   b. *I was never given the opportunity* to do what I liked. (∼ they never gave me the opportunity) <FRED CON008>

2. Prepositional dative structures were coded even if a corresponding double object structure would have been ungrammatical in standard English due to a pronominal direct object, as in (123-a). In a similar vein, some double object structures that would be ungrammatical in standard English, as in (123-b), were included in the tally:

   (123) a. Oh nono, he says, don't *give it to him*, he says . . . <FRED HEB018>

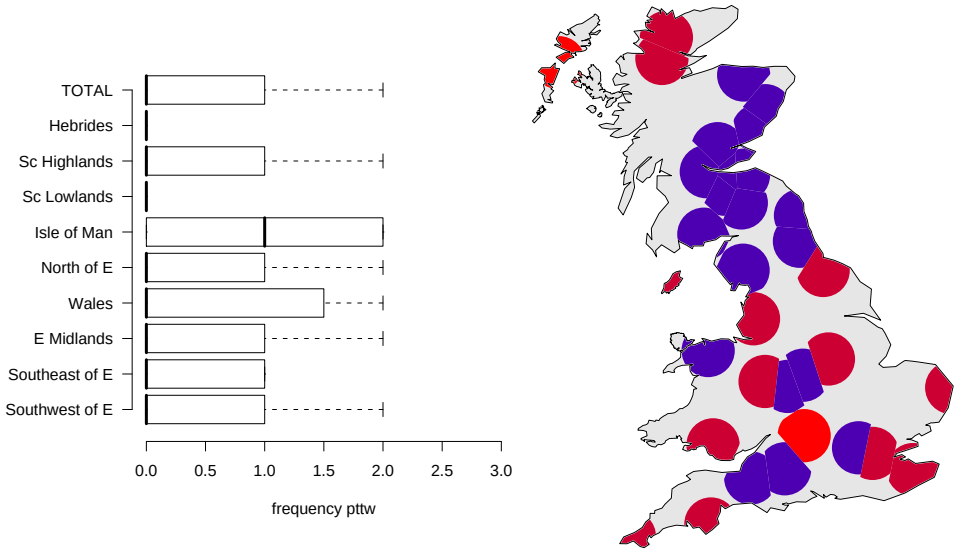

Figure 56: Feature [56] (prepositional dative constructions after GIVE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

b. ... but he was on leave and he *give me it* ... <FRED LAN010>

3. Direct and indirect objects, as well as the noun heads of prepositional datives, could be relativized, as in (124):

(124) That was the first nail in his coffin, the way he had scraped up and saved all his life, and we done without lots of things, [*that* he could *have been giving us*] ... <FRED DUR003>

4. Direct objects were permitted to be instantiated by a nominal clause, as in (125):

(125) And he used to generally *give him* [*what it cost*], so he didn't lose too much money. <FRED KEN002>

Finally, a retrieval script identified and registered all manually tagged prepositional dative and double object constructions after the verb GIVE.

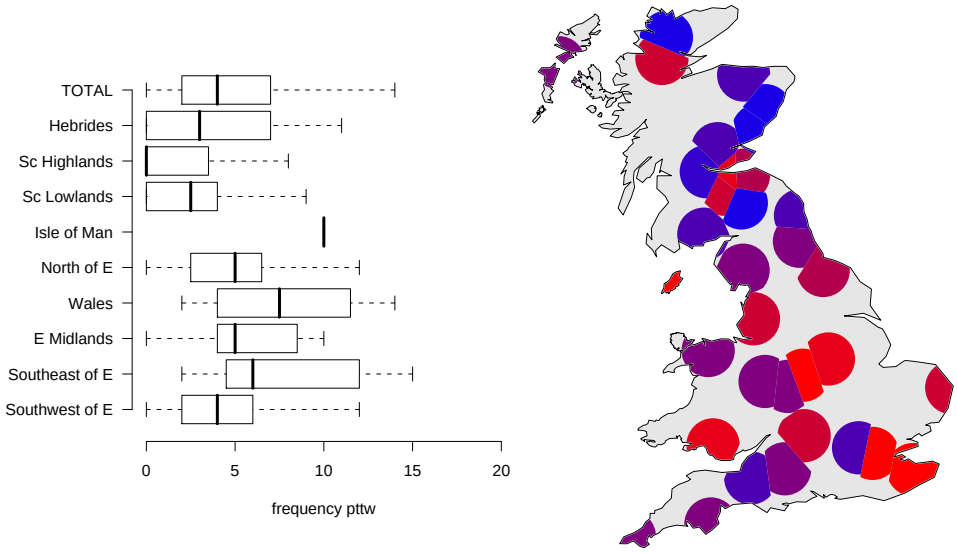

Figure 57: Feature [57] (double object constructions after GIVE). Left: variance by *a-priori* dialect area. Right: projection of relative frequencies to geography.

6. Summary statistics

Table 1 provides a number of summary statistics. For each individual feature, we report:

- which dataset was utilized ($\text{FRED}_{\text{full}}$ versus $\text{FRED}_{\text{abridged}}$) (cf. Section 4.2);
- the number of raw hits of the feature in this dataset;
- the resulting mean normalized frequency *pttw* of the feature;
- as a summary measure of FRED-internal dispersion, the standard deviation associated with the normalized frequency figure, for both the county level ($N = 34$ objects) and the location level ($N = 156$ objects).

feature	dataset	raw hits	mean frequency *pttw*	st.dev. county level	st.dev. location level
[1]	full	158	0.6	0.6	0.9
[2]	full	1,282	5.2	2.5	4.7
[3]	full	185	0.8	1.0	7.8
[4]	full	440	1.8	8.7	28.1
[5]	full	3,302	13.5	6.8	9.4
[6]	full	14,721	60.1	21.9	32.1
[7]	full	109	0.4	0.4	1.1
[8]	full	1,404	5.7	4.4	6.8
[9]	full	1,115	4.6	3.6	4.5
[10]	full	853	3.5	1.7	2.8
[11]	full	1145	4.7	2.8	6.9
[12]	full	407	1.7	2.0	3.5
[13]	full	25,231	103.0	27.0	46.9
[14]	full	90,778	370.5	62.9	93.1
[15]	full	31,829	129.9	20.9	41.3
[16]	full	412	1.7	2.0	3.0
[17]	full	628	2.6	1.7	3.1
[18]	full	4,155	17.0	8.3	10.8
[19]	abridged	2,119	42.2	23.4	42.4
[20]	abridged	3,857	76.8	54.6	62.1
[21]	abridged	1,648	32.8	16.9	36.1
[22]	full	503	2.1	2.6	1.4
[23]	abridged	1,571	31.3	15.8	28.6
[24]	full	821	3.4	2.8	4.8
[25]	full	6,553	26.7	10.4	16.5
[26]	full	1,571	6.4	3.5	6.8
[27]	full	339	1.4	1.8	1.0
[28]	full	262	1.1	1.4	2.1
[29]	full	649	2.6	2.6	3.1
[30]	full	663	2.7	2.5	2.8
[31]	full	648	2.6	27.3	15.3
[32]	full	205	0.8	0.7	1.7
[33]	full	1,170	4.8	3.1	5.5
[34]	full	5,260	21.5	9.7	15.7
[35]	full	946	3.9	4.2	7.5
[36]	full	2,225	9.1	4.0	6.2
[37]	full	2,400	9.8	4.8	8.7
[38]	full	971	4.0	2.3	4.8
[39]	full	3,234	13.2	12.2	14.0
[40]	full	122	0.5	0.4	1.0
[41]	full	192	0.8	1.6	2.8
[42]	full	1,834	7.5	4.0	6.9
[43]	full	134	0.5	0.5	1.1
[44]	abridged	435	8.7	8.0	14.1
[45]	abridged	304	6.1	20.8	16.1
[46]	abridged	734	14.6	18.6	19.1
[47]	abridged	149	3.0	2.8	4.9
[48]	abridged	787	15.7	9.1	16.8
[49]	full	241	1.0	0.8	1.6
[50]	full	175	0.7	1.4	2.2
[51]	full	379	1.5	1.4	1.8
[52]	full	563	2.3	1.8	3.7
[53]	full	5,125	20.9	9.4	12.8
[54]	full	506	2.1	2.1	4.1
[55]	full	330	1.3	2.4	2.0
[56]	full	143	0.6	0.7	1.4
[57]	full	1,571	6.4	2.8	4.2

Table 1: The 57-feature catalogue: some summary statistics.

References

Anderwald, L. (2009). *The Morphology of English Dialects*. Cambridge: Cambridge University Press.

Aston, G. and L. Burnard (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bauer, L. (1994). *Watching English change: an introduction to the study of linguistic change in standard Englishes in the twentieth century*. London: Longman.

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

Comrie, B. (1976). *Aspect: an introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press.

Hernández, N. (2006). *User's Guide to FRED*. `http://www.freidok.uni-freiburg.de/ volltexte/2489/`. Freiburg: English Dialects Research Group.

Hinrichs, L. and B. Szmrecsanyi (2007). Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics 11*(3), 437–474.

Hundt, M. (2004). Animacy, agentivity, and the spread of the progressive in Modern English. *English Language and Linguistics 8*(1), 47–69.

Kortmann, B. and B. Szmrecsanyi (2004). Global synopsis: morphological and syntactic variation in English. In B. Kortmann, E. Schneider, K. Burridge, R. Mesthrie, and C. Upton (Eds.), *A Handbook of Varieties of English*, Volume 2, pp. 1142–1202. Berlin/New York: Mouton de Gruyter.

Leech, G. and J. Culpeper (1997). The Comparison of Adjectives in Recent British English. In T. Nevalainen and L. Kahlas-Tarkka (Eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, pp. 125–132. Amsterdam: Rodopi.

Mondorf, B. (2003). Support for *more*-support. In G. Rohdenburg and B. Mondorf (Eds.), *Determinants of Grammatical Variation in English*, Topics in English Linguistics, pp. 251–304. Berlin, New York: Mouton de Gruyter.

Orton, H., S. Sanderson, and J. D. A. Widdowson (1978). *The Linguistic Atlas of England*. London, Atlantic Highlands, N.J.: Croom Helm.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London, New York: Longman.

Sanderson, S. and J. Widdowson (1985). Linguistic geography in England: Progress and prospects. In J. M. Kirk, S. Sanderson, and J. D. A. Widdowson (Eds.), *Studies in linguistic geography: The dialects of English in Britain and Ireland*, pp. 34–50. London: Croom Helm.

Szmrecsanyi, B. (2005). Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory 1*(1), 113–150.

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English: a corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin, New York: Mouton de Gruyter.

Szmrecsanyi, B. (2008). Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing 2*(1–2), 279–296.

Szmrecsanyi, B. (2010). Geography is overrated. In S. Hansen, C. Schwarz, P. Stoeckle, and T. Streck (Eds.), *Dialectological and Folk Dialectological Concepts of Space*. Berlin: Walter de Gruyter.

Szmrecsanyi, B. (in preparation). *Morphosyntactic Variation in British English Dialects: Exploring Dialect Grammars in the Aggregate Perspective*.

Szmrecsanyi, B. (submitted a). Aggregate data analysis in variationist linguistics.

Szmrecsanyi, B. (submitted b). Corpus-based dialectometry – a methodological sketch.

Szmrecsanyi, B. and N. Hernández (2007). *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. `http://www.freidok.uni-freiburg.de/volltexte/2859/`. Freiburg: English Dialects Research Group.

Szmrecsanyi, B. and L. Hinrichs (2008). Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. In T. Nevalainen, I. Taavitsainen, P. Pahta, and M. Korhonen (Eds.), *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, pp. 291–309. Amsterdam: Benjamins.

Tagliamonte, S. and H. Lawrence (2000). "I Used to Dance, but I Don't Dance Now": The Habitual Past in English. *Journal of English Linguistics 28*, 324–353.

Trudgill, P. (1999). *The dialects of England* (2nd ed.). Cambridge, MA, Oxford: Blackwell.

Viereck, W., H. Ramisch, H. Händler, P. Hoffmann, and W. Putschke (1991). *The computer developed linguistic atlas of England*. Tübingen: Niemeyer.

Voronoi, G. (1907). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik 133*, 97–178.