Awareness of a syntactic change in Shetland

November 26, 2015

1 Quantifying speakers' impressions of syntactic changes

In this work we investigate the human capacity for tracking changes in syntactic variables by probing speakers' awareness of three instances of the loss of verb movement in the variety of Scots spoken in Shetland. (Some more information on Shetland/Shetland Scots here.) The changing variables in question are:

- verb positioning in imperatives: change from verb inversion (VS) to Standard SVO structure
- yes/no question syntax: change from VS with initial main verb to a 'periphrastic do' structure
- wh question syntax: change from WhVS with main verb to a 'periphrastic do' structure

In all three cases the usage is moving from quite frequent use of the incoming variant among the average speaker towards its near-categorical use in younger speakers (?)

1.1 Methodology

To quantify people's explicit knowledge about ongoing language changes we adapted a self-evaluation method originally used to investigate the perception of phonetic changes by **?** and **?**, who asked speakers to self-assess their relative usage of several phonetic variables. We refined the methodology, so that every sociolinguistic variable under investigation was covered by a one page questionnaire eliciting both speakers' estimates of their own usage, as well as that of other social groups, and even properties of the variants themselves. At the top of each questionnaire page, the two competing syntactic variants were introduced in the following way:

> You are probably familiar with these two ways of asking somebody to do something:
> *"Mak du dy ain denner!"* *"Du mak dy ain denner!"*

The order of the two variants was randomised between individuals, in the above example the outgoing variant is on the left, the incoming one (akin to Standard English "You make your own dinner!") on the right. The dialectal spelling of the example sentences is quasi-standardised on Shetland, and their mixing with the Standard English formulations of the questionnaire is not unusual. The actual questionnaire consisted of the following five questions which tapped into different aspects of people's explicit knowledge about the changes in question:

Question 1: "How much do you use either of these variants?"
This explicit question regarding speakers' own frequency of use could be answered on a 5-point scale, with the options labelled 'I use only (V1)', 'I use mostly (V1)', 'I use both equally', 'I use mostly (V2)' and 'I use only (V2)', with the order of V1 and V2 matching those of the presentation of the two variants above.

Question 2: "How much do you think are people around you using either of the variants?"
This question could again be answered on a 5-point scale with options 'People use only (V1/2)/mostly (V1/2)/both equally'. This question does not just provide information on speakers' perception of their average interlocutors' frequency of use, but the *relative difference* between the answers to questions 1 and 2 can potentially provide information on whether speakers think of themselves as being 'ahead' or 'behind' the curve of a particular change relative to their speech community.

Question 3: "Which of the two variants do you think is *older*?"
This (intentionally vague) question is intended to get at speakers' beliefs or connotations regarding the 'age' of the competing variants, without yet drawing explicit attention to the fact that the variable is in fact changing. The three possible answers were 'V1 is older', 'V2 is older' and 'People have always used both', with the order of V1 and V2 randomised.

Questions 4+5: "How much do you think *younger/older speakers* use either of the variants?"
The final two questions tap into speakers' awareness of the apparent time development of a change, with the same 5-point options as above: 'Younger/older speakers use only (V1/2)/mostly (V1/2)/both equally'. The order of the two questions was randomised between individuals.

Data collection proceeded in three stages: first, to pilot the methodology, 8 participants were asked to complete the paper version of the questionnaire on site in Shetland for two variables with the following example sentences:

1. verb positioning in imperatives: *Mak du dy ain denner!* vs. *Du mak dy ain denner!*, with the latter (incoming) variant akin to Standard English syntax, i.e. 'You (sg.) make your (sg.) own dinner!'

2. negation marking: *He didna go* vs. *He didnoo go* – this stable variable was added as a control, with 'didna' being the widespread variant against very localised used of 'didnoo' on the island of Whalsay to the East of Shetland's main island (more explanation?)

Following the successful pilot, 16 more participants were asked to complete an extended 4-page version of the questionnaire which covered two further variables:

3. yes/no question syntax: "Kens du Sarah?" vs. "Does du ken Sarah?", 'ken' being the Scots lexeme for 'to know'

4. wh question syntax: "Whit gae du him?" vs. "Whit did du gie him?"

These first 24 participants were part of a balanced sample matched for gender, age, and geographic location within Shetland. All participants grew up in Shetland, were currently living in Shetland, and hadn't lived outside Shetland for more than X years (typically to study at university before returning). In all cases, the questionnaire was administered as an exit-questionnaire following a ~40(?) minute task which involved providing grammaticality judgments for a large number of examples of the changing variables in question (as well as fillers?), which was carried out in pairs.

Finally, we created an identical online version of the 4-variable questionnaire which was advertised via email and social networks. The online questionnaire was self-contained (i.e. not preceded by the grammaticality judgment task) and provided us with a convenience sample of another 53 participants from all over Shetland. Apart from their age, gender and current geographical location we also collected information on all participants' occupation, where in Shetland they grew up, any extended times they spent outside the isles, as well as the origin of their parents.

1.2 Results

Pooling together the data from the paper-based and online questionnaires, the total number of responses was $N = 77$ for the imperative and negation variables, and $N = 69$ for the yes/no as well as wh question syntax. We will go through the results question by question.

1.2.1 Demographic information of the sample

Both the locally collected and online samples had a similar age distribution, with participants ranging from 18 to 73 years old with a mean age of around 40. Report on geographical distribution and socioeconomic background?

1.2.2 Self-estimates of own usage

People's assessment of their own usage is traditionally not regarded as reliable since self-reports often reflect a communities' overt prestige values rather than people's actual usage (Labov, Trudgill, inter alia). Assuming the self-reports *were* accurate in our case, we might expect individual responses to be predicted well by the speakers' age, with younger speakers claiming higher usage levels of the incoming variants. While the distribution of ages per response shown in Figure 1 suggests that there might be such an effect the amount of data per response is strongly skewed, with the majority of responses falling onto just three options of our 5-point scale. Using a conditional inference framework to

construct a recursive binary partitioning (?) with R's `party` package[1] we find that, from all possible predictor variables, it is actually the *sociolinguistic variable* itself that best predicts the distribution of responses. The partitioning in Figure 2 shows that for both wh and yes/no questions all but one of participants report using the incoming variant at least half the time, while responses for the imperative show a much flatter distribution (this result is confirmed by an ordinal logistic regression model using the response distribution of the imperatives as a baseline, with $p < .0001$). This first result indicates that the change in verb position in imperatives might be lagging behind the two question variables. This conclusion receives independent support from the grammaticality judgments elicited from the first 24 participants, which exhibited high acceptability for both imperative variants, in contrast to comparatively lower acceptability ratings for the outgoing question forms.

Beyond the role of sociolinguistic variable, conditional inference also finds a second significant binary partitioning for the two question types, with the participants' age a good predictor of their self-evaluation responses (left branch in Figure 2). The effect is in the direction we would expect, with younger participants mostly claiming categorical usage of the incoming variant, while older speakers are most likely to still report some usage of the outgoing question variants. This result is borne out beyond the strict binary split by an ordinal logistic regression model with age as a continuous predictor variable, at significance level $p < .01$).

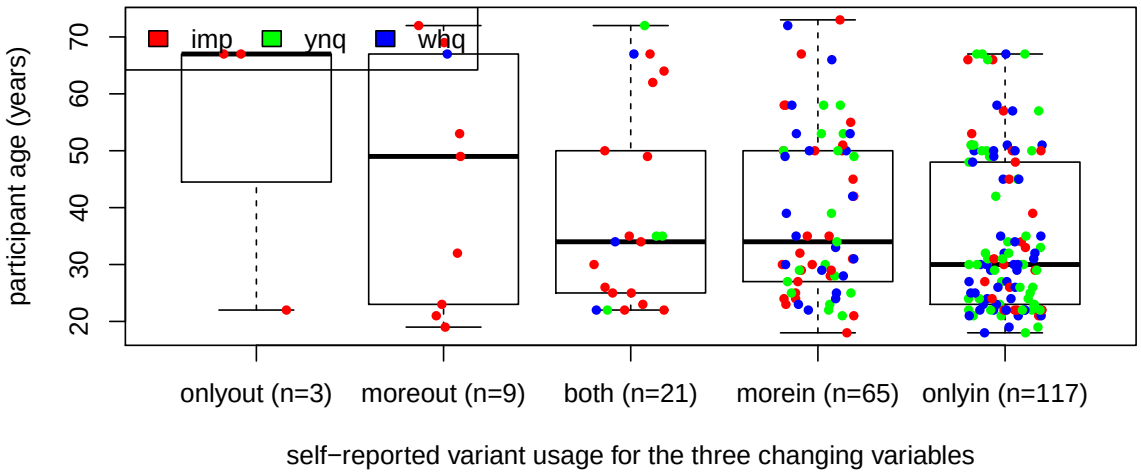

Figure 1: Age distribution per self-use response for all three changing variables pooled together. While all of the five possible responses were selected by people across all ages, the visualisation suggests that younger people are more likely to self-report higher usage of the incoming variants.

[1] `https://cran.r-project.org/web/packages/party/`

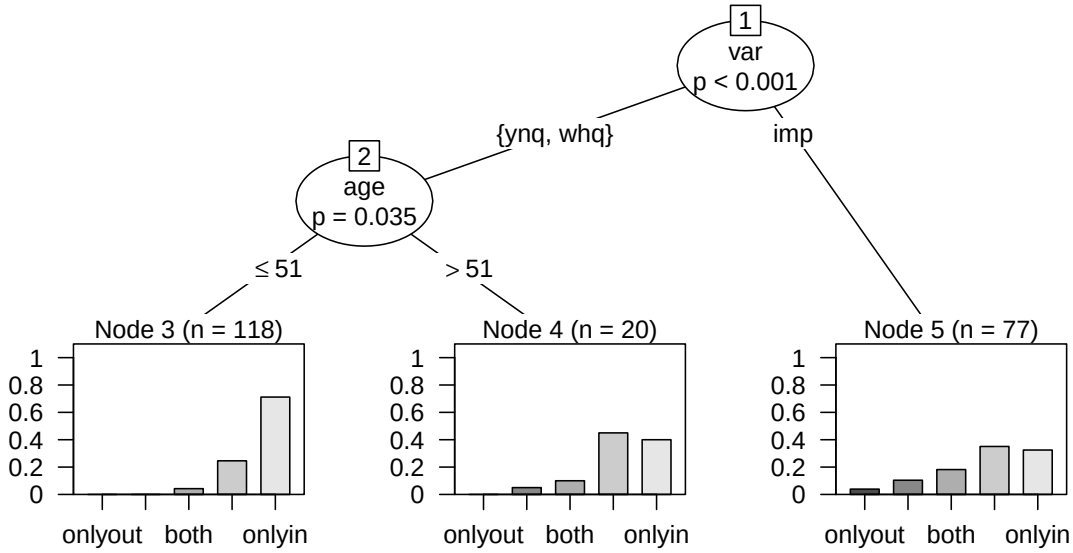

Figure 2: Distribution of the self-reported usage levels for the three changing variables along the 5-point scale. Imperatives exhibit higher self-reports of using the outgoing variants (rightmost node), while most people claim categorical or near-categorical use of the incoming variant for both question types. Within the responses for the two question variables (left branch), age is a significant predictor of a binary partitioning with a cutoff age of 51, with younger speakers most likely to report using only the incoming variant, and older speakers more likely to select 'mostly' the incoming variant.

(optional TODO? Correlate responses of the first 24 participants with the grammaticality judgments obtained from them for cross-validation – gotta select proper subset of judgment stimuli that matches the awareness questionnaires in terms of verb arity etc.)

1.2.3 'Other people' usage estimates

When it comes to reporting on the linguistic usage levels of other individuals in their speech community the overall pattern is similar to the self-evaluation responses, but with an added central tendency or edge-avoiding effect in the responses, as can be seen in Figure 3. This presumably stems from the fact that, when imagining an 'average' individual, the informants will model this on the population average, which is almost necessarily non-categorical.

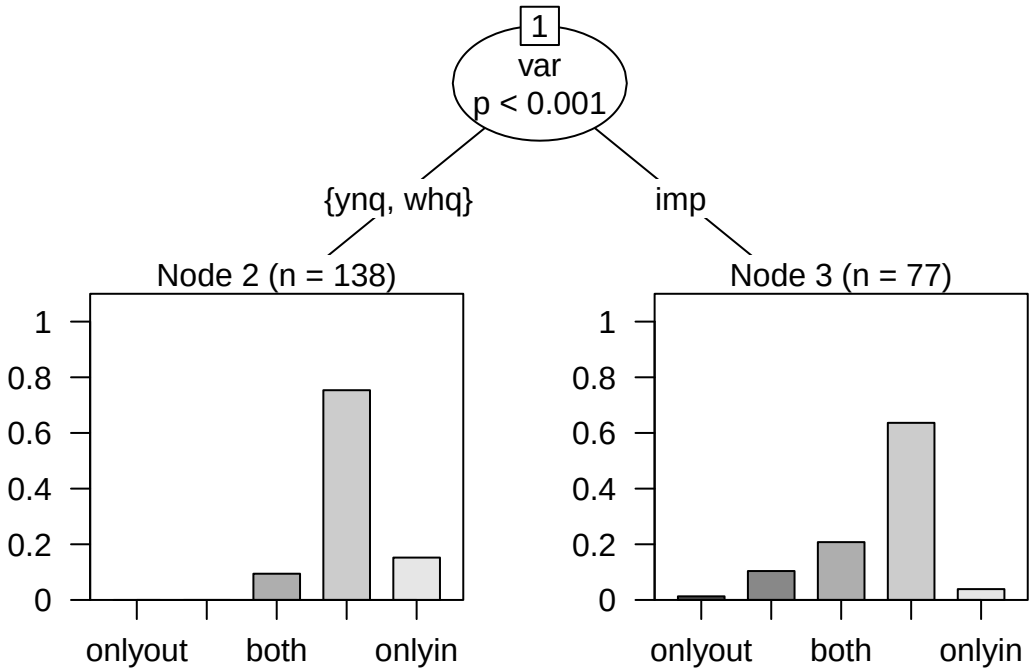

Figure 3: Speakers' estimates of the population-level usage of the three changing variables is best predicted by the sociolinguistic variable, with the two question types again clustering together and the imperatives showing a more spread-out distribution of responses. Both distributions are similar to the self-evaluation responses shown in Figure 2 except that they are shifted away from both of the extreme options, indicating that the population average is perceived to be variable rather than categorical.

The data from the first two questions implicitly contains another interesting piece of information, namely where the speakers regard their own variable usage to be relative to the community-level. We measure this by looking at the number of ordinal categories along the 5-point scale that separates the self-evaluation vs reported community-level usage, where positive numbers indicate that a speaker reported a relatively higher usage of the incoming variant for themselves than for the community. The participants' age is a good predictor of this difference between themselves and the community, as can be seen in Figure 4. The binary partitioning suggested by conditional inference shows that the majority of under-58-year-olds regard themselves as 'ahead' of their community in terms of using the incoming variants, whereas those over 58 are most likely to report being level with the community (an ordinal regression model using age as a continuous predictor variable is significant at p=0.0022545).

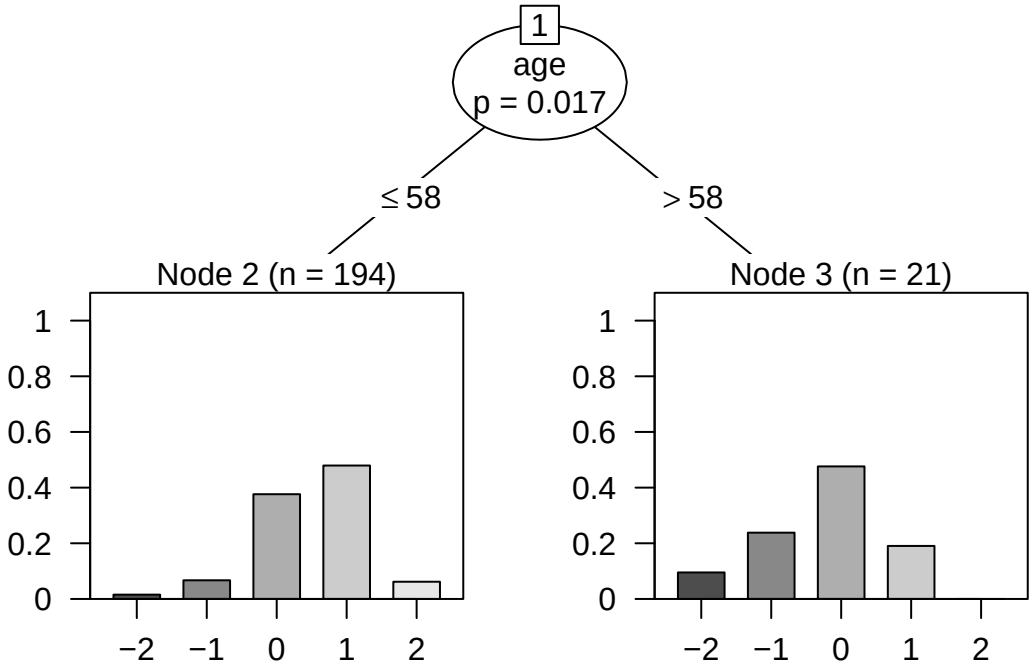

Figure 4: Conditional inference tree predicting the relative difference between the self-evaluation usage level vs. reported community usage level for the three changing variables, in number of ordinal categories on the 5-point scale. Positive numbers indicate that a speaker reported a higher usage of the incoming variant for themselves than for the community, and vice versa. No speaker indicated their own usage to be more than two ordinal categories away from the community level. Younger speakers are more likely to perceive themselves to be ahead of the community level usage, while older speakers are most likely to report their usage to be level with the community.

1.2.4 Beliefs about the age of competing linguistic variants

The third question of the questionnaire aimed at eliciting the speakers' beliefs about the variants by explicitly asking which they thought was the 'older' of the two, with 'people have always used both' given as a neutral third option. Results show that, for the three changing variables, people reliably identify the outgoing variant as the 'older' one. For the stable negation control variable results are more mixed, but many also report the less widespread 'didna' variant as being older. Conditional inference on the data shows that the changing vs. stable variant division is the only significant predictor of the participants' responses to the question (see Figure 5). While this result indicates that the community shares common beliefs about the directionality of these changes, these beliefs could be based on any or all of real time observation of the change, connotations of variants being archaic, or apparent time differents in variable usage across age groups (although up to this point in the questionnaire it was avoided to draw explicit attention to such differences).

Previous sociolinguistic research which has revealed gender differences where women were often found to be leading linguistic changes has been used to argue that, due to their

social position, women are more sensitive to sociolinguistic cues (Labov, inter alia). Interestingly, for the three changing variables pooled together, whether participants correctly report the outgoing variant to be older is not predicted well by any of the participant variables, in particular not by gender or age ($p > .1$, logistic regression model).

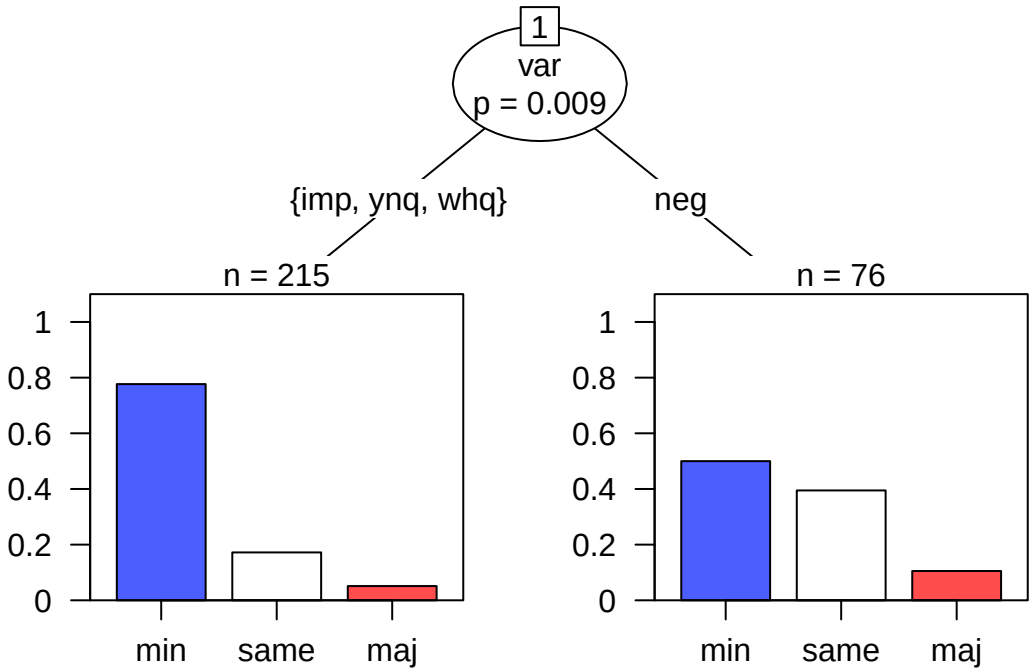

Figure 5: The type of sociolinguistic variable is the only significant predictor of the participants' response to the question "Which of the two variants do you think is *older*?". There appears to be no effect of age, gender, order of presentation or whether the questionnaire was filled out on-site following the grammaticality judgment task vs. online. For the three changing variables (left branch) most individuals report the outgoing forms which have already become the *min*ority variants to be older. For the stable negation variable most respondents also picked the minority (because geographically limited) 'didna' variant, but with more answers falling on the majority variant as well as the "people have always used both" option.

1.2.5 Perception of apparent time differences

The final pair of questions, which ask for the participants' impression of the relative usage level of the two competing variants in otherwise underspecified 'younger' and 'older speakers' was aimed at helping us to determine whether individuals can perceive and report apparent time differences in categorical variables. As can be seen in Figure 6, speakers consistently report higher usage of the majority incoming variant among younger speakers.

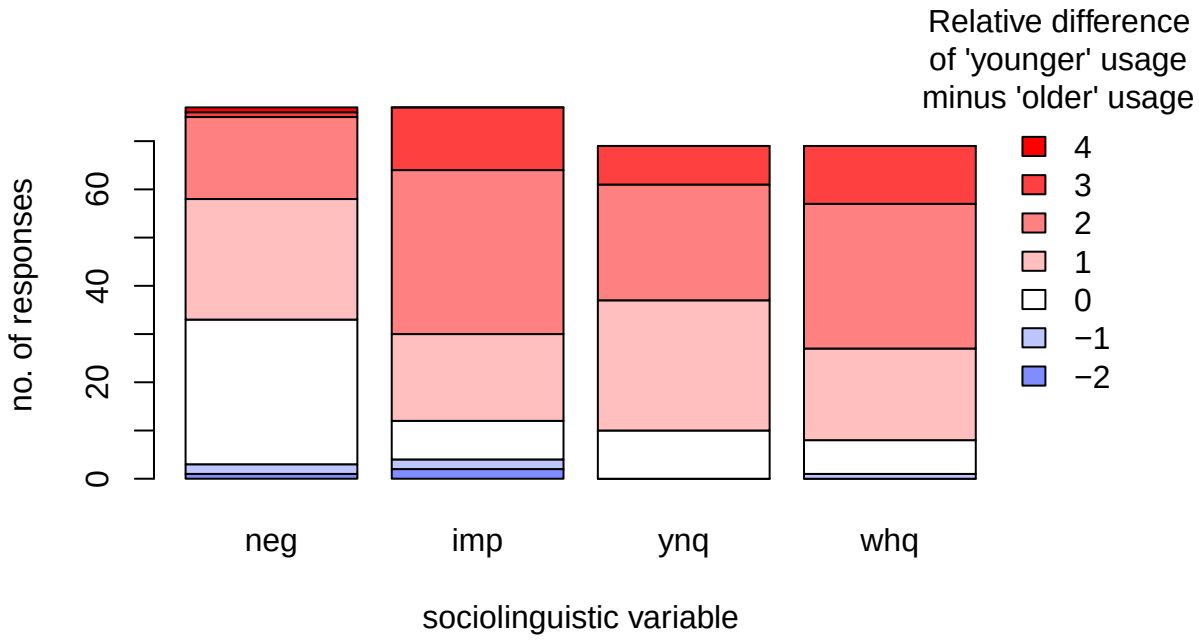

Figure 6: Relative difference between the reported usage of the two variants between 'older' and 'younger' speakers, for each of the four variables. The relative difference indicates the number of ordinal categories that separates an individuals' responses for the two age groups along the 5-point scale, where positive numbers (in shades of pink) correspond to reporting higher usage of the majority variant among younger speakers, and vice versa. For the three changing variables (on the right) this majority variant is indeed the incoming variant.

While this might suggest that individuals can accurately perceive and report on apparent age differences in variable usage, we cannot straightforwardly jump to this conclusion. The conditional inference model in Figure 7 shows that our participants' responses to question 3, and in particular the 'wrong' classification of the incoming variant as the 'older' one, is a significant predictor of whether younger speakers are reported to be 'ahead', 'behind' or 'level' in their relative usage of the incoming variant. Conversely, the direction of the relative ordinal difference between the reported younger/older speaker usage is also a significant predictor of the answer to the preceding 'Which of the two variants is *older*?' question, shown in Figure 8, with a reported apparent time difference (i.e. an ordinal difference > 0) predicting increased identification of the outgoing variant as the 'older' one.

Based on this it remains an open question whether people might have inferred their answer to question 3 based on an apparent time difference they perceived, or if participants felt led to answer questions 4 and 5 in a way that would post-hoc justify their response to question 3, which might have been based on other (socio-indexical) knowledge. One potential way to find out whether the latter is the case is by checking whether the absolute progression of the changes, where the imperatives seem to lag behind a bit, is also evident in the answers about the specific age group questions. The overall pattern of

relative apparent age differences does not differ for the questions vs. the imperatives (ordered regression model predicting the relative difference between responses to the last two questions, from -4 to 4, with sociolinguistic variable as a predictor, $p = 0.2495938$). The *absolute* magnitude of the answers to both the younger and older speaker groups does however differ for imperatives, with generally lower usage levels reported for the imperatives ($p = 1.4109191 \times 10^{-4}$), as we would expect.

To avoid the same problem, an improved methodology should therefore randomise the order of questions 3 vs. 4+5 between individuals, and ideally also embed 4+5 in a bigger set of distractor questions.

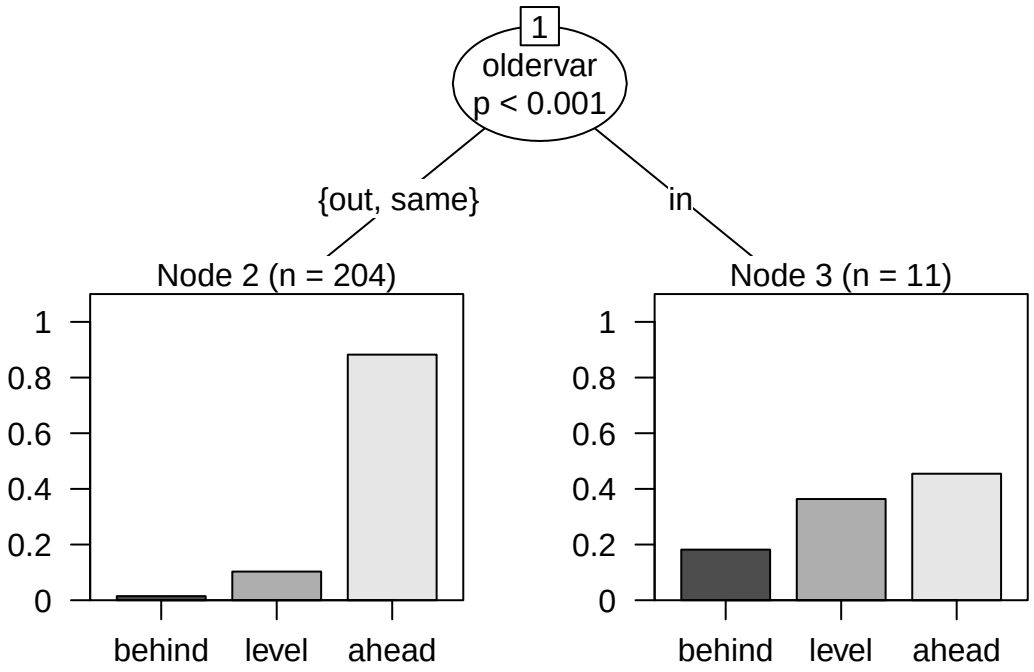

Figure 7: Predicting individual apparent time differences of the three changing variables: whether the incoming variant is incorrectly reported as being the 'older' of the two variants (right branch) is a significant predictor of whether younger speakers are reported to be 'ahead', 'behind' or 'level' in their relative usage of the incoming variant. This result could indicate that answers about the age of the variants were based on perceived differences in the apparent time distribution, but it could also be that the responses to the later apparent time questions were influenced (so as to provide a post-hoc justification of the earlier variant age answer).

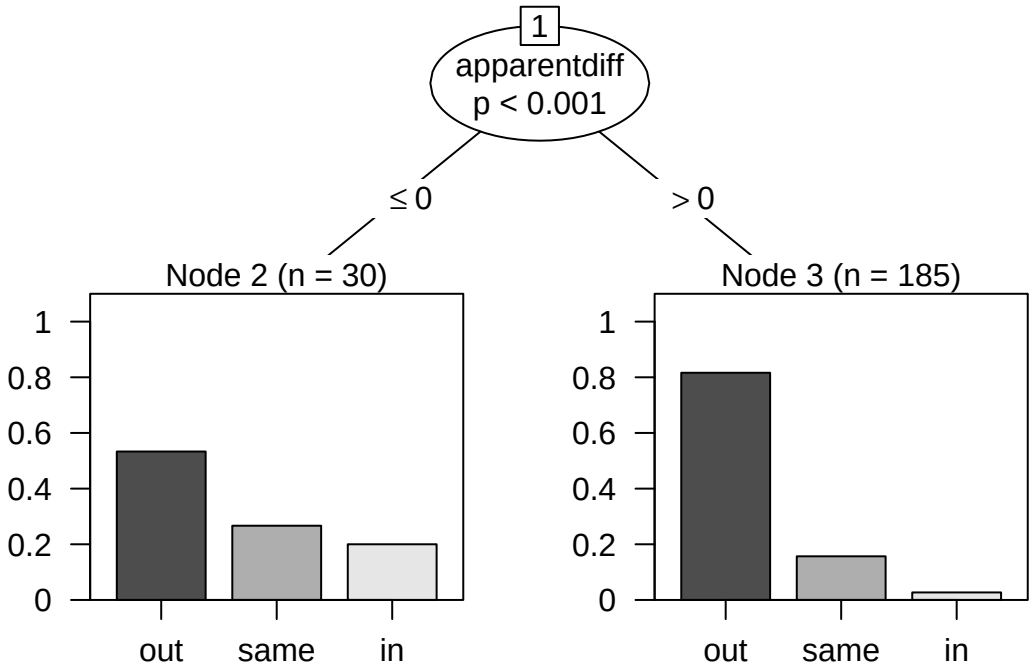

Figure 8: Using reported apparent time differences of the three changing variables as a predictor: the direction of the relative ordinal difference between reported younger/older speaker usage (`apparentdiff`) is a significant predictor of the answer to the preceding 'Which of the two variants is *older*?' question, with reports of younger speakers' increased use of the incoming variant predicting improved identification of the outgoing variant.

1.3 Using variant age responses to control for leading questions in apparent time responses

```
##
##        -2 -1  0  1  2  3  4
##   out   0  1 15 48 72 31  0
##   same  0  2  6 15 12  2  0
##   in    2  0  4  1  4  0  0
```

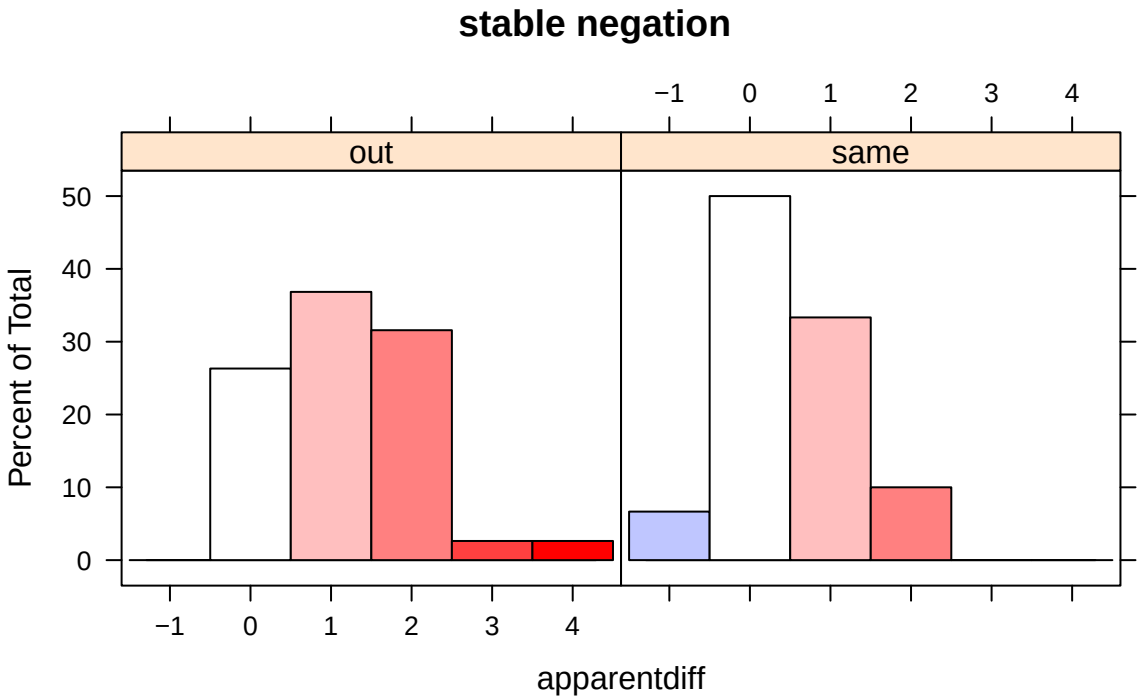

Figure 9: Reported apparent time difference per answer to the 'which variant do you think is older' question. While at least some of the apparent time responses for the changing variables go away when the participants did not report having any age beliefs, the apparent time reports disappear much more strongly for the negation, indicating that the reported differences may have been driven by the preceding question only. Still outstanding is a test against the underlying baseline distributions given the responses, which are plotted below

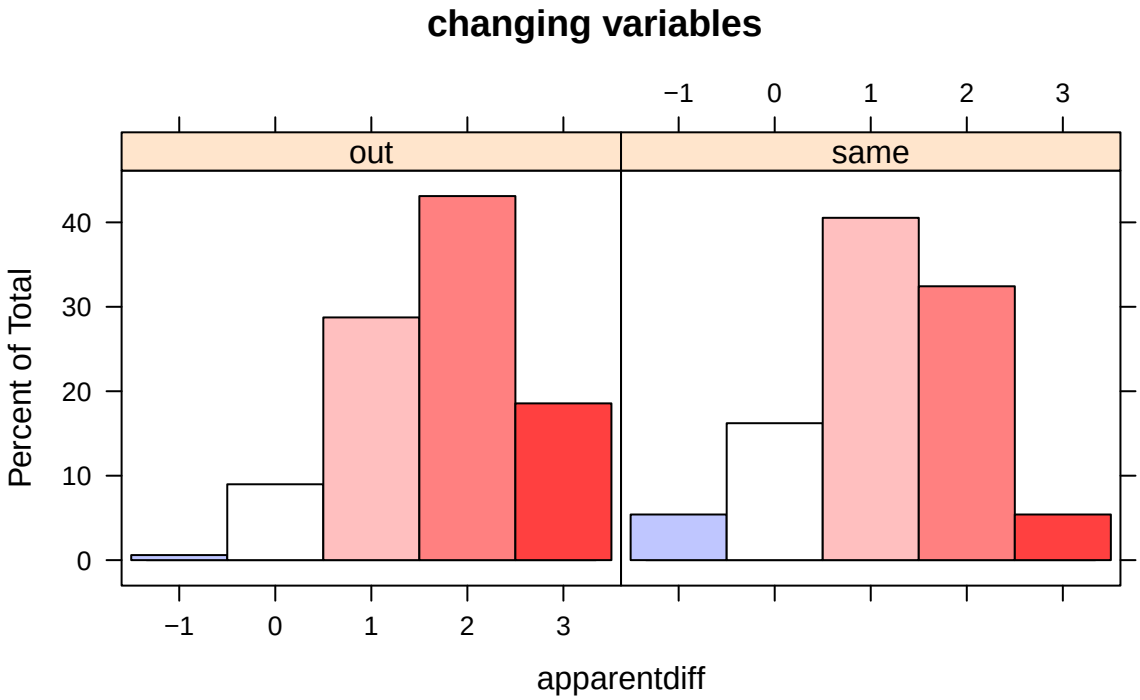

Figure 10: Reported apparent time difference per answer to the 'which variant do you think is older' question. While at least some of the apparent time responses for the changing variables go away when the participants did not report having any age beliefs, the apparent time reports disappear much more strongly for the negation, indicating that the reported differences may have been driven by the preceding question only. Still outstanding is a test against the underlying baseline distributions given the responses, which are plotted below

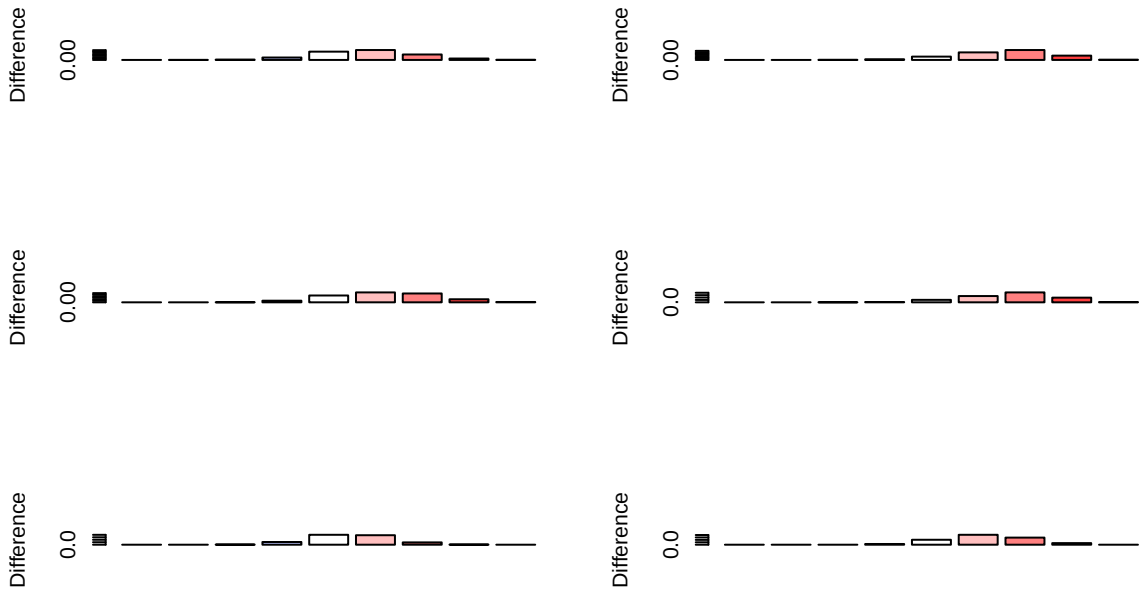

Figure 11: Reported apparent time difference per answer to the 'which variant do you think is older' question. While at least some of the apparent time responses for the changing variables go away when the participants did not report having any age beliefs, the apparent time reports disappear much more strongly for the negation, indicating that the reported differences may have been driven by the preceding question only. Still outstanding is a test against the underlying baseline distributions given the responses, which are plotted below

2 Summary of relevant results per predictor

Following the ad-hoc approach of determining the best predictors for every answer via binary partitioning as done above, we can probe the data for particular effects that we would have expected.

2.1 Gender effects

The only significant gender effect that can be found is in the first two questions, i.e. the reported 'self' as well as 'rest of the population' usage levels, where females tend to report *higher* levels of the incoming variant both for themselves (for all four variables) as well as the general population (only for the changing variables). While not significant for the four variables taken individually the effect goes in the same direction in all cases, with it being strongest for the two question types, somewhat less for the imperatives, and unreliable/inconsistent for the negation variable. The bare data of the responses split by variable and gender is shown in Figure 13. This gender effect does *not* translate to a higher difference between perceived self-usage and perceived community usage, i.e. females do

not perceive themselves to be further 'ahead' the community than males do, the answers to both questions appear to be shifted in unison. Also, despite an often purported increased sensitivity to linguistic changes among females, neither the identification of the 'older' variants nor the strength of the perceived apparent time differences show any effects of gender.

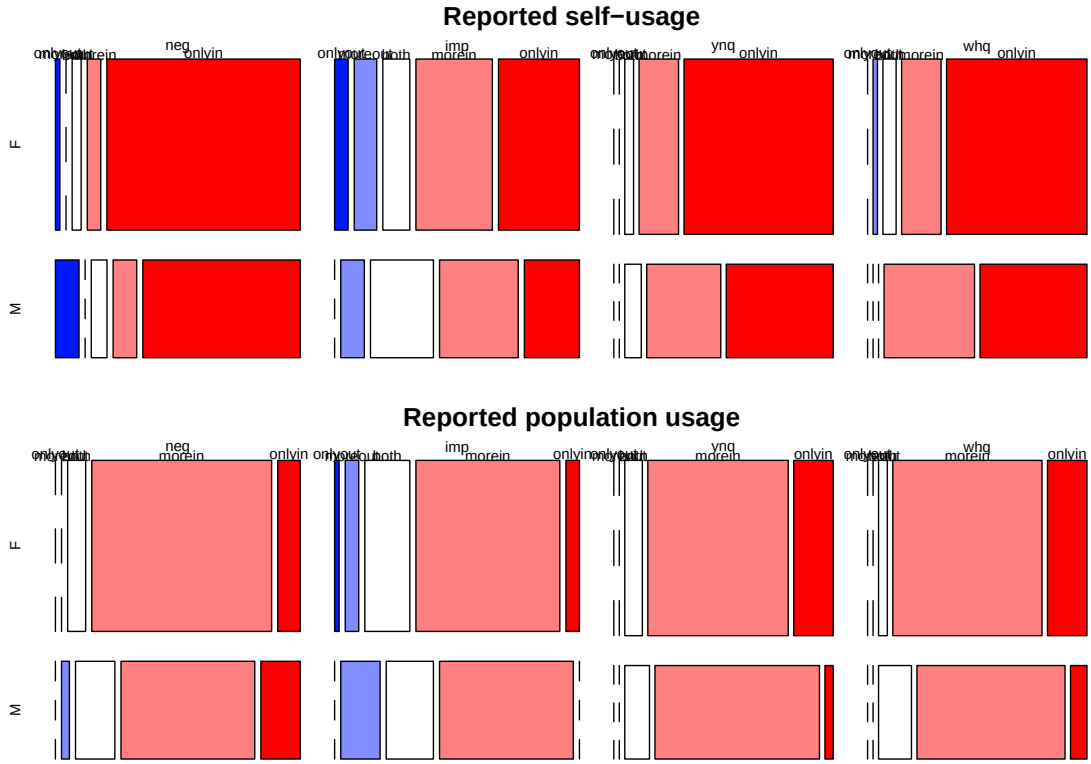

Figure 12: Raw data for responses regarding self-usage (left) and general population/people around you usage (right), split by sociolinguistic variable (x-axis) and gender (y-axis). For the changing variables, females generally report a higher level usage of the incoming variable (redder responses) for both self-usage as well as perceived community average than do males. Higher bars for females than males indicate the skewed distribution towards female participants in the convenience sample.

2.2 Age effects

There are several significant age effects: firstly, age is a good predictor for participants' reported self-use of the changing variables, in particular for the dataset of responses to the wh questions, as could be seen in Figure 1. Combined with no age effect for participants' estimates of the community-level usage, age is consequently a good predictor for how much the respondents think they are 'ahead' or 'behind' their estimated community level usage, as discussed in section 1.2.3. The effect size for these two effects is in the range of -0.0265, meaning that for every year somebody is older, their probability of choosing the next-lower category increases by about 2.5% (the coefficient is given in log-odds, so the cumulative effect is not linear but slightly less so). While the effect of age on self-reports

could already be seen in Figure 1, a visualisation of this effect on the difference between reported self and community-level usage is shown in Figure 14.

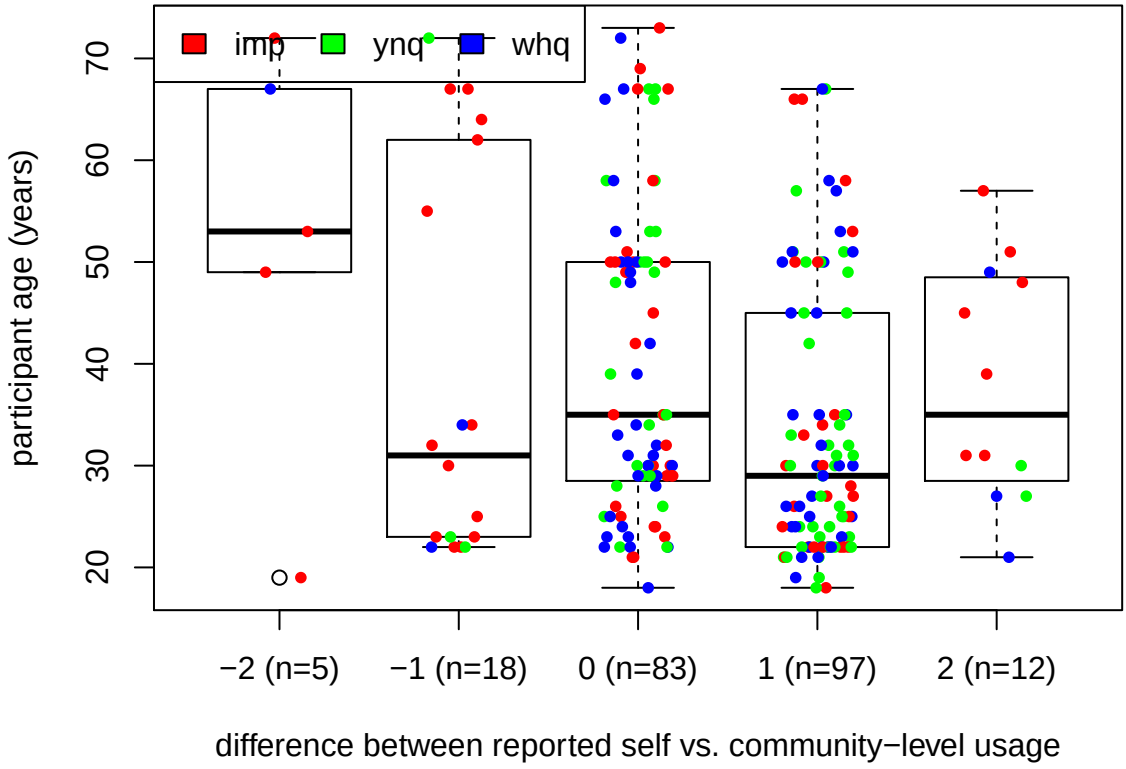

Figure 13: Age distribution of the difference between reported self and community-level usage. Apart from the moderately visible age effect, it is intriguing that very few individuals report themselves to be 'behind' the curve for the question types, as opposed to the imperatives. This effect might be a matter of the measure used here: the imperatives, still very much a change in progress, allow for a wider range of putting yourself 'relative' to the community usage (on either extreme). Maybe, instead of measuring the absolute difference between selected categories per individual, the relative difference should be measured relative to the distribution of community-level responses?

For the changing variables, there is no significant age effect on perceived 'age' of the variants, the reported apparent time differences, or on the estimates of how much 'younger' and 'older' speaker groups use the variants.

2.2.1 Age effects on the stable negation variable

Since age does not affect the reported self-usage levels of the negation variable, and the reported community-level usage is consistent for all variables, there is consequently no

effect of age on how much people report to use the negation variants relative to their surrounding community.

In terms of perceived age of the variants, the only significant age effect is actually with the stable negation variable, where older people are more likely to identify the geographically more widespread 'didnoo' variant as being the 'older' one (caveat: that is still only 8 out of 76 people, and those 8 happen to contain relatively older speakers).

The other question is whether apparent time negation answers are predicted by user age – while there is no significant effect for the changing variables, age is a borderline significant predictor for the apparent time reports for negation. However, this effect goes away once the answer to question 3 is taken into account, i.e. which variant the informants thought were older explains most of their reported apparent time differences. The raw data for informant age vs. reported apparent time differences (with notes on the statistical test results) is shown in Figure 15.

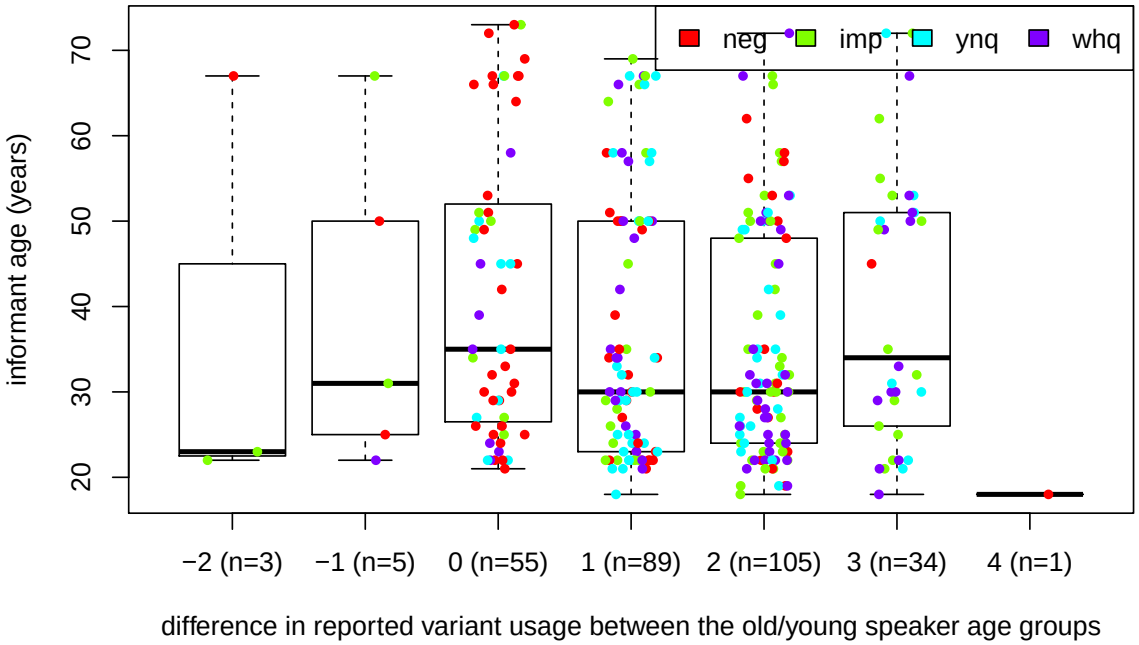

Figure 14: Age distribution of the difference between reported 'older' and 'younger' speaker usage levels for all four sociolinguistic variables. Response '0' means that the informant reported the same usage levels for both the 'older' and 'younger' speaker groups, reponse '1' means that younger speakers were reported to be one ordinal category more advanced in their usage of the incoming variant, etc. It is evident that the 0 response works as a 'hard edge' (hardly anyone reports older speakers to be more advanced than younger ones), which goes against the assumptions of the ordinal regression method used. As a consequence, re-levelling the responses (or running a different statistical technique) could potentially yield a much bigger effect of age.