2018-03

III: ANALYSES AND RESULTS FOR STUDY 1: ESTIMATING THE EFFECT OF LINGUISTIC DISTANCE ON VOCABULARY DEVELOPMENT.

Floccia, Caroline

10.1111/mono.12350
Monographs of the Society for Research in Child Development
Wiley

Chapter 3 – Analyses and Results for Study 1: Estimating the Effect of Linguistic Distance on Vocabulary Development

Caroline Floccia, Thomas D. Sambrook, Claire Delle Luche, Rosa Kwok, Jeremy Goslin, Laurence White, Allegra Cattani, Emily Sullivan, Kirsten Abbot-Smith, Andrea Krott, Debbie Mills, Caroline Rowland, Judit Gervain, Kim Plunkett

Corresponding author: Caroline Floccia, caroline.floccia@plymouth.ac.uk

To estimate the impact of linguistic distance on vocabulary development, which was the aim of Study 1, we needed to account for the effects of all situational factors that were known or suspected to shape bilingual development. This was achieved through a two-step analyses of the data from the 372 children whose Additional Language was one of our 13 target languages.

Plan of Analyses

In the first step, analyses were conducted on variables already established within the literature as strong predictors of vocabulary size (relative amount of exposure to each language in child-directed speech and overheard speech, gender and SES). Analyses were conducted initially in ANCOVAs (to include continuous variables such as LEQ) and then subjected to confirmation in linear mixed models, with variables entered as fixed effects predictors only if they reached significance in the ANCOVAs.

In the second step, analyses were then conducted on less well-established variables (factors relative to source, properties and status, and measures of linguistic distance), in models containing verified predictors from the initial stage. Again, ANCOVAs were followed by linear mixed models with the same logic as in the first step.

The reasons behind this two-step process extend beyond the aims of Study 1, as explained below. The effects from the ANCOVAs would hold for a population with the same breakdown of language communities as those in our sample (e.g., which was 17.4% German, 12.8% French etc.). The linear mixed models do not make that assumption and are thus strongly

preferred if conclusions are to be generalised to all bilinguals. The purpose of the linear mixed models was indeed to provide the best possible test for the importance of candidate predictor variables for bilinguals generally, not merely those whose Additional Language was one of those 13 used in this paper (and therefore preparing for Study 3). For this it was imperative that Language Community be modelled as a random, not a fixed effect, as the ANCOVAs would do.

In contrast, the value of the ANCOVAs is that they allow a straightforward test of the significance of simultaneously entered predictor variables, something that is problematic with linear mixed models due to a lack of consensus on how to compute the degrees of freedom for each predictor (Baayen, Davidson, & Bates, 2008). Our procedure was thus to perform a preliminary selection of significant predictors from ANCOVAs, subject to confirmation in linear mixed models, before finally being included in predictive models.

Following steps 1 and 2, predictive linear mixed models of expressive and receptive vocabulary for the UKBTAT tool (Study 2) were calculated with predictors retained only if their effect size in the ANCOVAs was larger than $\eta^2 = 0.02$ (Cohen, 1988), which is considered a threshold for small effects in ANOVAs and multiple regressions. These final models do not include measures of linguistic distance, since we aimed at developing norms which could be applied to any bilingual children learning British English, and measures of linguistic distance would not be available for Additional Languages that are not amongst our 13 target Additional Languages. Predictive models for the UKBTAT (Study 2) and the test of their generalisation beyond the 13 target Additional Languages (Study 3) are presented in Chapter 4.

Predictor Variables

Step 1 predictors were language exposure scores (LEQ and Overheard speech), Gender, and SES. Language Community (which Additional Language is spoken by the child) was included as a 13 level dummy variable simply to control for its effects. Step 2 predictors were

the mode of exposure variables, and three Linguistic Distance variables. Mode of exposure variables belonged to three categories: *source* of each language (whether two parents are native Additional Language speakers or only one, total number of English speakers, total number of Additional Language speakers, number of siblings, time spent in English speaking daycare, time spent in Additional Language speaking daycare), *properties of the input* (degree of language use consistency in parents' speech, proportion of native English, proportion of native Additional Language), and *status of the Additional Language* (Welsh group vs. all other Additional Languages). Linguistic Distance variables were phonological overlap, word order typology and morphological complexity.

Test language, or TestLang (English/Additional Language), was included as a repeated measures factor to examine the differential impact of predictors on English and Additional Language. See Table 1 for a summary of these variables.

Vocabulary Measures

Dependent variables were CDI counts of receptive and expressive vocabulary in English and the Additional Language. We conducted analyses on two different versions of the CDIs. Our starting point was a 30 word CDI made up of those 30 words present in the Oxford Short Form CDI and all 13 Additional Language CDIs (see Table A5-1 in appendix 5). This 30-word CDI had the advantage of holding items constant across a child's two test languages, thus controlling, amongst other things, for word frequency (although frequencies between translation equivalents differed, correlations were in the order of .8 over the 13 English-Additional Language pairings). The disadvantage of this approach is that, of the words common across all CDIs, a disproportionate number were high frequency words. This results in a ceiling effect with, for example, over a third of children scoring 100% on English comprehension. Our second CDI used the full 100 words of the Oxford Short Form and the full Additional Language CDI re-represented as a percentage, to accommodate the fact that the standardised CDIs varied

considerably in length for each language. While these data suffer from no ceiling effects and maximise the sampling of vocabulary, they are the least satisfactory in that there is no control of word frequency between all Additional Language CDIs, that is, obtaining a 40% score in the German CDI is not necessarily equivalent to a 40% in the Portuguese CDI. The main analyses reported here were performed on the 30-word CDI, with a specific section added for the data from the full CDIs. Given that the results were essentially similar when using the 30-word CDI or the full CDIs, please note that the final equations in the UKBTAT (Chapter 4) are calculated from the 100 words of the Oxford Short Form for English (for increasing representativeness of the model's coefficients), and the 30 words for Additional Languages when relevant.

Descriptive Statistics

Predictors. Tables 2 to 6 presented summary data of all predictor variables broken down by Language Community. Because of the strong associations between Language Community and predictor variables, it was important to use models that included Language Community to avoid attributing to predictor variables the explanatory power that was actually simply due to variability over language communities.

Vocabulary measures. All data for vocabulary measures in English (for the full cohort) are reported in Table 7, and in Table 8 for each Additional Language (for the 372 target Additional Languages learners only). On the Oxford Short Form CDI, children understood on average 67.9 words and produced 41.2 words (variables CDI100Comp and CDI100Prod). On the Additional Language CDIs, which varied in length, children overall understood 54.9% of Additional Language words and produced 24.2% (PropALcomp and PropALprod). When restricting the analysis to the 30 words common to all CDIs, children understood on average 24.4 English words and produced 17.0 (variables CDI30comp and CDI30prod). For the target Additional Language, children understood on average 21.7 words and produced 11.2

(ALCDI30Comp and ALCDI30Prod), which was significantly less than in English (comprehension: paired t (371) = 6.25, $p < .0001$; production t (371) = 11.51, $p < .0001$). As noted before, all main analyses provided below were run on the 30-word CDIs, with analyses on the full form CDIs provided in a specific section.

Of interest is the comparison of the bilingual scores to monolinguals. Based on the Oxford CDI database for 125 monolingual children aged 23.0 to 25.0, 24-month-olds understand 73.6% of the 416-word CDI (SD = 16.9) and produce 48.3% (SD = 25.8). To compare these scores to those of the Oxford Short Form CDI, we applied a correction ratio computed from the comparison between the long and short CDI described in the "Measuring vocabulary" section (see Methods). A score in the long CDI divided by 0.90 provides an equivalent score on the short CDI. That means that the 24-month-old monolinguals are estimated to understand 81.8% of words and produce 53.7% if assessed with the Oxford Short Form CDI. In contrast, the cohort of 430 bilinguals understood 67.9% of the Oxford Short Form CDI (SD = 25.0) and produced 41.2% (SD = 26.0), which is significantly less than the monolinguals (comprehension: t (553) = 5.84, p = .0001; production: t (553) = 4.74, p = .0001).

INSERT TABLE 7

INSERT TABLE 8

Results

Step 1 - Predictors firmly established in the literature. We looked at four predictors: LEQ (proportion of English in child-directed speech), Overheard speech (proportion of English spoken between parents), SES and Gender. Three indices (income, maternal education and

paternal education) were initially selected as potential predictors for SES, but due to high correlations between income and parental education, and because income had the widest observed range, the latter was selected. Results are very similar if parental education is used. Thus our step 1 ANCOVA consisted of the between-subjects predictors of LEQ, Overheard speech, Income and Gender, with the within-subjects predictor of TestLang (Additional Language/English). Language Community was included as a between-subjects factor, since we wished to ascertain in Study 3 (as aforementioned) the degree to which all other predictors are generalisable to bilingual 24-month-olds regardless of the particular Additional Language she or he is learning. Separate ANCOVAs were run for production and comprehension scores. All ANCOVAs used as dependent variables the 30-words CDIs (see Table 9 and 10 for full results in comprehension and production respectively).

INSERT TABLE 9 HERE

INSERT TABLE 10 HERE

For comprehension there was no main effect of LEQ, but an interaction of LEQ and TestLang ($F_{1,342}$ = 75.07, p < .001, η^2 =.18). Analysis of the effect of LEQ on each test language separately revealed that it significantly reduced Additional Language scores ($F_{1,342}$ = 18.81, p < .001, η^2 =.05) and increased English scores ($F_{1,343}$ = 21.46, p < .001, η^2 =.06): the more child-directed English children heard, the more English they understood, and the less Additional Language they understood (see Figure 1). Overheard speech (the proportion of English vs the Additional Language spoken between the parents when the child was present) significantly increased comprehension scores overall ($F_{1,342}$ = 10.55, p = .001, η^2 =.03), and showed an interaction with TestLang ($F_{1,342}$ = 38.72, p < .001, η^2 =.10). Breaking down the effect of

Overheard speech for the two test languages, while no effect was seen for Additional Language ($F < 1$), a beneficial effect was seen for English ($F_{1,342} = 32.42$, $p < .001$, $\eta^2 = .09$). The more English spoken between the parents, the more beneficial effect on English comprehension (see Figure 2). A main effect of Income was found ($F_{1,342} = 11.97$, $p = .001$, $\eta^2 = .03$) and no interaction with TestLang. There was no main effect of Gender ($F_{1,342} = 1.39$, $p = .24$) or interaction with TestLang ($F_{1,342} = .05$, $p = .82$).

INSERT FIGURE 1 HERE

For production no main effect of LEQ was observed but again an interaction with TestLang was seen ($F_{1,342} = 91.58$, $p < .001$, $\eta^2 = .21$). As in comprehension, individual analyses on each of the test languages revealed a negative effect of LEQ on Additional Language scores ($F_{1,342} = 17.09$, $p < .001$, $\eta^2 = .05$) and a positive effect of LEQ on English scores ($F_{1,342} = 25.94$, $p < .001$, $\eta^2 = .07$) (see Figure 1). Overheard speech showed no main effect but did show an interaction with TestLang ($F_{1,342} = 34.08$, $p < .001$, $\eta^2 = .09$). Breaking down the effect for the two test languages showed a significant detrimental effect on Additional Language ($F_{1,342} = 3.91$, $p = .049$, $\eta^2 = .01$) and a beneficial effect on English ($F_{1,342} = 143.30$, $p < .001$, $\eta^2 = .04$) (see Figure 2).

INSERT FIGURE 2 HERE

Income did not have a significant effect ($F_{1,342} = 3.39$, $p = .07$) and there was no interaction with TestLang ($F_{1,342} = .55$, $p = .46$). There was a main effect of Gender ($F_{1,342} = 21.00$, $p < .001$, $\eta^2 = .06$), with girls outperforming boys and no interaction with TestLang ($F_{1,342} = .14$, $p <= .71$) (see Figure 3).

INSERT FIGURE 3 HERE

Linear mixed models were then carried out with fixed effects only for those predictors that reached significance in the aforementioned ANCOVAs, with random slopes and intercept for Language Community and random intercepts for participants. Separate models were conducted for each fixed effect variable, with significance assessed by comparing each model against a null in which the fixed effect was absent. Linear mixed models were calculated using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in the R environment (R Development Core Team, 2006, version 0.99.896). The coefficients for each model are given in Tables 11 (comprehension) and 12 (production). Note that the effect that is tested in these comparisons is the combined main effect and interaction with TestLang if one was indicated by the ANCOVAs.

INSERT TABLE 11

INSERT TABLE 12

For comprehension there was a significant effect of LEQ ($\chi^2(2) = 18.02$, $p < .001$) and Overheard speech ($\chi^2(2) = 18.62$, $p < .001$) but no effect of Income ($\chi^2(1) = 1.60$, $p = .21$). For production there was a significant effect of LEQ ($\chi^2(2) = 18.75$, $p < .001$), Overheard speech ($\chi^2(2) = 23.59$, $p < .001$) and Gender ($\chi^2(1) = 13.26$, $p < .001$). When these analyses were conducted again with each significant fixed effect entered into a model already containing the other significant fixed effects (Table 13), the last entered fixed effect retained its significance ($p < .006$) in each case. Because we consider the linear mixed models to be the more appropriate

significance test for a model that generalises to all bilinguals (Study 3), Income was discarded as a predictor for subsequent analyses. The remaining predictors at the end of Step 1 were the relative amount of exposure to English in child-directed speech (LEQ), the proportion of English in parental overheard speech (Overheard speech), and Gender. This first step allowed us to confirm the robustness of predictors from the literature for the building of a model of the child's lexicon (Study 2), based on data from the 13 target Additional Languages learners.

INSERT TABLE 13 HERE

Step 2 - Secondary predictors. Secondary variables were then added to ANCOVAs containing those predictors shown to be significant in the Step 1 analysis above (LEQ, Overheard speech and Gender), with the 30 words common to all CDIs as dependent variables. These predictors were all the mode of exposure variables, and the three Linguistic Distance variables.

Mode of exposure variables were assessed by adding them individually to ANCOVAs containing LEQ, Overheard speech and, in the case of production, Gender. Societal status (variable Status), degree of language use consistency in parents' input (variable Consistency), presence of siblings (variable Siblings), and number of parental native Additional Language speakers (one or two; variable FamLang) were added to a model containing TestLang as a within–subjects factor. Models with only Additional Language or English test scores omitted the factor of TestLang but included factors describing the native input of test language (variables PropEngN and PropALN), the number of sources of test language (SourcesEng and SourcesAL), and the amount of day care provided in the test language (EngDaycare and ALDaycare).

Only two variables achieved significance: Consistency and PropEngN (see table A7-1 in appendix 7). Consistency interacted significantly with TestLang in determining production scores ($F_{1,355} = 3.94$, $p =. 047$, $\eta^2 = .01$), due to English vocabulary being boosted by a decreasing consistency in parents' use of the two languages ($F_{1,355} = 6.07$, $p = .014$, $\eta^2 = .017$): the more parents used a mix of English and the Additional Language, and the more English vocabulary was produced. The proportion of parental native English spoken (PropEngN) significantly improved English production scores ($F_{1,296} = 4.12$, $p = .043$, $\eta^2 = .01$).

INSERT TABLE 14 HERE

The effect of Consistency on English production scores was confirmed with a linear mixed model ($\chi^2(1) = 5.79$, $p = .016$), however, owing to the very small effect size, this variable was not included in the UKBTAT predictive models. The effect of proportion of native English spoken on English production scores was not supported by a linear mixed model ($\chi^2(1) < 1$), and was not retained in the UKBTAT equations.

Finally, the three Linguistic Distance variables were assessed, phonological overlap, word order typology and morphological complexity. Because these showed a perfect to very high association with Language Community, these factors could not be added to ANCOVAs and were assessed in linear mixed models only (see Tables 15 to 17). These revealed a significant effect of Phonological Overlap on Additional Language production ($\chi^2(1) = 4.61$, $p = .032$), a significant effect of Word Order typology on Additional Language comprehension ($\chi^2(1) = 6.02$, $p = .014$), and a significant effect of Morphological Complexity on Additional Language comprehension ($\chi^2(1) = 4.80$, $p = .028$). In all three cases, an advantage was found for children learning an Additional Language close to English. No effects on English were seen.

INSERT TABLE 15 HERE

INSERT TABLE 16 HERE

INSERT TABLE 17 HERE

In summary, all variables but two (Consistency and PropEngN) from Step 2 analyses were excluded at the ANCOVA stage, due to lack of significance. Consistency did have a significant effect in the subsequent linear mixed models but its effect size was too small to warrant integration in the UKBTAT predictive models (Study 2, next Chapter). PropEngN (proportion of English that is native) failed to reach significance in the linear mixed models, and therefore will not be included in the UKBTAT models. All measures of Linguistic Distance were found to be significant in the linear mixed models, fulfilling the predictions of Study 1. It must be kept in mind however that all results obtained at Step 2 should be taken in the context of multiple comparisons and viewed as subject to confirmation in future studies.

Metrics of linguistic distance will not be included in the UKBTAT models (Study 2) because our aim was to build a predictive model for any bilingual child learning British English (Study 3), and measures of linguistic distance will not be available for any Additional Language different from our 13 target languages.

Comparison with full CDI. The effects found in the Step 1 analysis were checked in the full CDI data (the proportion of words in the 100-word Oxford Short Form CDI and in the original Additional Language CDIs). The pattern of significance was identical with effect sizes highly comparable. In particular, linear mixed models once again showed no effect of Income on comprehension ($\chi^2(1) = 1.36$, $p = .24$).

On a final note, in the analyses provided in this paper, we have deliberately ignored item-level analyses as they are beyond our current scope (but see Table A5-2 in appendix 5 for a breakdown of comprehension and production of each English word in the 30-word CDI, as a function of exposure). However, further analyses at this level would provide a privileged insight of the internal organisation of the bilingual lexicon, complementing pioneering studies regarding the processes by which new words are added to the lexicon in monolinguals (e.g., Hills, Maouene, Maouene, Sheya, & Smith, 2009) or bilinguals (Bilson, Yoshida, Tran, Woods, & Hills, 2015). By comparison, where Bilson et al. (2015) collected data in 181 children spanning the age of 6 months to 78 months from eight different English-Additional Language pairs, using a version of the MCDI Toddler form designed for children aged 16 to 30 months of age (Fenson et al., 2007), our dataset includes data for both English and Additional Language (when available) from 430 24-month-olds, collected using age-appropriate tools. One application of this data now currently being conducted is to examine whether phonological overlap modulates the 2-year-old bilingual lexicon in terms of associative relationships and frequency for translation equivalents.

Key Findings

The aim of Study 1 was to establish whether measures of linguistic distance could predict vocabulary size at age 2 in bilingual toddlers. To achieve this, we first showed that vocabulary size was modulated by a set of robust factors from the literature: relative amount of exposure to English versus the Additional Language (English and Additional Language comprehension and production), proportion of English in overheard speech (English comprehension and production) and gender (English and Additional Language production). Two other, less established predictors emerged: first, English production was boosted when each parent used a mix of languages as compared to using one language only – but due to its very small effect size this variable will not be included in the subsequent UKBTAT equations.

Second, the proportion of native English spoken by parents significantly improved English production scores – but this effect did not survive in linear mixed models and therefore will not be included in the UKBTAT. No other variables reached significance, in particular not SES.

Importantly, we found that the three measures of linguistic distance predicted vocabulary knowledge in two-year-olds learning British English and one of 13 target Additional Languages, above and beyond variance due to Language Community: phonological overlap, SVO order typology and morphological complexity. Specifically, we found that children learning a language phonologically close to English had a larger production vocabulary in their Additional Language; similarly, those learning typologically or morphologically close languages to English had a larger Additional Language comprehension vocabulary.