Chapter 2 – Methods

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

The methods described in this chapter cover the cohort data collection, which constituted the common data set for all three studies reported in this paper. A sample of 430 bilingual toddlers, learning British English and one of 13 target Additional Languages (N = 372), or any other Additional Language (N = 58), were identified over a 2-year period. To increase variation in both English/Additional Language pairs and in the situational factors outlined above (language exposure, mode of exposure and demographic factors), data was collected through trained research assistants recruited in the six universities involved in this project (Bangor, Birmingham, Kent, Liverpool, Oxford and Plymouth), as well as in Bristol and Leicester, each having access to multilingual populations to various degrees. However, since the testing platform was remotely accessible, the final sample comprised families from all areas of the UK, apart from Scotland and Northern Ireland. Bangor had the additional advantage of being located in a region with 75% bilingual Welsh-English children, providing a unique opportunity to compare language skills in bilinguals growing in a region with predominant bilingualism, to those whose bilingualism is linked to immigration.

When the child approached her second birthday, volunteer parents were contacted via the website UKBilingualToddlers, and the following data were collected in this order: English expressive and receptive vocabulary as measured through a bespoke Oxford Short Form CDI (Hamilton et al., 2000); Additional Language vocabulary as measured through the corresponding version of the CDI, when available; a family questionnaire with detailed questions about demographics (developed for the UK-CDI standardisation project, Alcock et

al., in prep); and the Plymouth Language Exposure Questionnaire (Cattani et al., 2014) which provided the LEQ measure of relative exposure to each language.

To sum up, the current study measured four parent-assessed outcome variables at 24 months: receptive English vocabulary, expressive English vocabulary, receptive Additional Language vocabulary and expressive Additional Language vocabulary (through CDIs). For each of these outcome variables, we investigated the influence of the following factors: 1) gender, 2) SES (as assessed via parental income and educational level), 3) proportion of child-directed speech in English (LEQ), 4) proportion of overheard parental speech in English, 5) factors related to the source of each language (whether two parents were native Additional Language speakers or only one, number of sources of English, number of sources of the Additional Language, time in daycare in each language, number of older siblings), 6) factors related to the properties of the input (degree of language use consistency in parents' input, number of native and non-native speakers in each language), 7) status of the Additional Language (societal vs. minority), 8) the particular language community (i.e. which of the 13 additional languages the child was exposed to) and 9) the linguistic distance between English and the Additional Language as measured by a) phonological distance, b) morphological distance and c) syntactic distance (see Table 1 for a summary of these variables).

INSERT TABLE 1 HERE

Method

Participants. Data were collected for a total of 430 children between February 2014 and July 2016. The data of an additional 31 children were discarded as they had hearing problems (N = 7), had a diagnosed developmental delay (as reported by parents; N = 6), were too young or too old (N = 17), or had incomplete records (N = 1). The data of another 41

children could not be included as parents did not complete the study. Out of the remaining final sample of 430 children (aged 23.89 months, SD 0.39, from 23.0 to 25.0; 193 girls and 237 boys), 372 were learning English and one of the 13 target Additional Languages: Bengali, Cantonese, Dutch, French, German, Greek, Hindi/Urdu, Italian, Mandarin, Polish, Portuguese, Spanish, and Welsh. Following King (2001), spoken Hindi and Urdu were classified as two varieties of the same language. The remaining 58 were learning English and one non-target Additional Language (see Table 2). The proportion of children born in the UK was 94.1% for the 372 children learning a target Additional Language, and 93.1% for the 58 non-target Additional Language learners. Out of these 430 children, the information for family income (an optional field) was not supplied for 15 children (13 in the target language community and 2 in the non-target language community). See Table 2 for a full description of the sample.

INSERT TABLE 2 HERE

Procedure and instruments. The data collection was initiated when the children reached 23.5 months old. When signing the online consent form on the UKBT database, parents were notified that there were four tasks to complete for the study: none of these tasks involved testing the children, allowing for remote data collection. Specifically, the CDIs and the family questionnaire were completed on the online platform by the parents, and the Plymouth Language Exposure Questionnaire was completed by the research assistants during a final telephone interview. A paper copy of the questionnaires was sent to parents who were unable to access the internet. For some families who did not feel confident in English, the research assistants met with the parent(s) to help them go through the various questionnaires. When signing up, contact information and identification of the language(s) being spoken at home triggered the selection of the appropriate Additional Language CDI when available.

Metrics of linguistic distance. To create a toddler-centric representation of language distance, each of the 406 non-onomatopoeic words from the Oxford CDI (Hamilton et al., 2000), as well as their translation equivalents across the 13 target Additional Languages, were transcribed into broad phonological representations. These were produced by trained phoneticians, each of whom was a native speaker of the language they were asked to transcribe. Our metric of language distance was then calculated as the overlap between the phonological representation of a word in British English and its translation equivalent in the Additional Language. This overlap was based upon the Levenshtein distance, that is, the minimal number of insertions, deletions and translations that are required to get from the British English phonological representation to that of the Additional Language. To produce a proportional measure of overlap this distance was subtracted from the length of the longest phonological sequence in British English or Additional Language, and then divided by the same number. This produces a measure of phonological overlap for each word, between 0 (no overlap) and 1 (perfect cognate), that preserves sequence order and is proportional to the length of the word.

$$Overlap = \frac{Max(BE\ length,\ AL\ length) - Levenshtein\ distance}{Max(BE\ length,\ AL\ length)}$$

An example of a calculation for the British English word “lamp” and its Italian translation equivalent “lampada” is shown below:

BE lamp /l.æ.m.p/ : Sequence length = 4

Italian lampada /l.a.m.p.a.d.a/ : Sequence length = 7

Levenshtein distance (l.æ.m.p, l.a.m.p.a.d.a) = 4 (1 translation + 3 insertions)

$$Overlap = \frac{Max(4,7) - 4}{Max(4,7)} = 0.43$$

The language level phonological overlap between British English and each of the 13 Additional Languages is shown in Table 1, calculated as the average overlap across all 406 words.

For the measure of word order typology, the Additional Language was assigned a 1 if it had a VO order like British English, a 2 if it had a mixed VO/OV order, and a 3 for a OV order (see Table 1). Finally, morphological complexity was assessed on a 3 point scale, with analytic/isolating languages (Mandarin, Cantonese) being ranked closer to English (value 1), followed by fusional languages such as French and German (value 2) and agglutinative languages such as Hindi/Urdu and Bengali (value 3) (see Table 3). To illustrate, in analytic Mandarin number is not marked on nouns, as in天 *yī tiān* "one day", 三天 *sān tiān* (lit.) "three day". In fusional French, the verbal suffix relates to grammatical mood, tense, aspect, person and number, as in *mangeais* "ate" (indicative, past, imperfective, second person singular) and *mangerions* "would eat" (conditional, present, perfective, first person plural). In agglutinative Bengali, nominative case for the word "river" is *nodi,* and the accusative *nodike*.

INSERT TABLE 3

Collecting demographic data. Demographic data were collected through the family questionnaire developed by Alcock et al. (in prep). This contains questions regarding (i) the health and development of the child, (ii) the child's family history, (iii) parental information (e.g., parents' educational level, income and postcode), and (iv) childcare arrangements. Some of these questions were repeated in the Plymouth Language Exposure Questionnaire (see below), but we tolerated overlap in order to retain each questionnaire's integrity. Following Arriaga et al. (1998), we focused on household income and educational levels when measuring SES, as typical indices of SES are highly correlated. Income was divided in four bands (variable Income), and education was measured on a seven point band that correspond to English qualification classifications, from no qualifications to a postgraduate degree (variables

MumEd and DadEd; see appendix 2). Education was chosen as it is generally used as a proxy for SES (e.g., Bornstein, Hahn, Suwalsky, & Haynes, 2003; Fenson et al., 2007), and it is usually a better predictor of language development than income (e.g., Hoff, 2003); in addition we estimated that in the case of immigrant families, educational level might better reflect the child's learning environment than mere economic circumstances. The educational status of both parents was used since the correlation between these two predictors was not large (r = .29).

As with the Fenson et al. (2007) and Hamilton et al. (2000) studies, the current study had an under-representation of low SES children within our bilingual cohort. This may be representative of SES distribution across the national population of bilingual children: Dustmann and Frattini (2011), using a variety of large scale British and international sources collected between 1993 and 2009, observed that immigrant populations in the UK tend to leave the education system later and have higher wages than their native peers. It is also likely that this under-representation stems from sampling, with low SES bilingual families reluctant to take part in research, especially in cases when they are not confident in English.

***Evaluating amount of exposure to each language*.** The Plymouth Language Exposure Questionnaire (Cattani et al., 2014) was used to obtain the percentage of direct language exposure received by the child in English and the Additional Language in a typical week based on a unique 5 to 10 minute phone interview (variable LEQ). The questionnaire (available at http://www.psy.plymouth.ac.uk/leq/) requested information about (i) the average number of hours spent by the child in nursery/with a childminder in each language environment (variables EngDaycare and ALDaycare); (ii) the language(s) spoken by each parent at home and the relative frequency of use of the two languages (variables MumPropEng and DadPropEng, measured on a 5-point scale); (iii) the number of hours spent by the child alone with each parent; (iv) whether the parents spoke equally with their child when both parents present; and (v) the number of hours of the child's sleep in a typical day (to evaluate the number of possible

contact time during a week). The detail of these variables and calculations leading to the proportion of English vs the Additional Language in a typical week (variable LEQ) is found in appendix 3.

To obtain the proportion of English/Additional Language in overheard speech (variable referred to as Overheard speech), an added question (5-point scale) was inserted after the original Plymouth Language Exposure Questionnaire (see appendix 2). See Table 4 for a summary of the results per language group.

INSERT TABLE 4 HERE

Evaluating the mode of exposure (source, properties and status). Measures of the various factors underpinning the *source of each language* were derived from questions which were part of the initial sign-up sheet, the family questionnaire and the Plymouth Language Exposure Questionnaire (see Table 5). Straightforward measures based on individual questions were the identification of the type of family (binary score for two parents native Additional Language speakers or only one; variable FamLang), the number of hours per typical week in English or Additional Language daycare (EngDaycare and ALDaycare), and the number of older siblings living in the house (until the age of 18 years; variable Siblings). Regarding the number of speakers in each language, a score of 1 was given to each native speaker parent, each older sibling, and attendance to a form of daycare (variable SourcesEng and SourcesAL, with an observed range of 0 to 6; see appendix 2).

INSERT TABLE 5 HERE

Regarding the *properties of the input* (see Table 6), the degree of language use consistency from each parent was obtained through the questions in the Plymouth Language Exposure Questionnaire asking parents to quantify on a 5-point scale their relative use of English and Additional Language. Specifically, a parent would obtain a 1 for always speaking Additional Language, 2 for usually speaking Additional Language, 3 for English and Additional Language half of the time, 4 for usually speaking English and 5 for always speaking English (variable MumPropEng and DadPropEng). Then the degree of consistency would be recoded as a minimum of 1 if the answer to the above was a 1 or a 5; a 2 if the answer to the above was a 2 or a 4; and a maximum of 3 if the answer to the above was a 3 (variables MumConsistency and DadConsistency, averaged as Consistency).

The proportion of native/non-native speech produced by parents was calculated from the same question, in conjunction with whether the parent was a native speaker of Additional Language or not. That is, the number of hours spent with each parent during a typical week was calculated as: 168 (number of hours in a week) - total sleeping time - hours in daycare - hours alone with the other parent (variable A). Then, each parent's score on their respective PropEng variable (1 to 5) was re-expressed as a proportion from 0 to 1 (1 = 0, 2 = .25, 3 = .5, 4 = .75, 5 = 1) to obtain the proportion of English in their speech (variable B). The resulting amount of English in this parent's input was obtained by multiplying A by B. If this parent was a British English native speaker, then AB would correspond to the amount of native input, and if the parent was an Additional Language native speaker, AB would be the amount of non-native input. The final proportion of native English input across both parents (the variable PropEngN), was obtained by dividing the total amount of native English by the sum of native and non-native English. The proportion of native Additional Language input (the variable PropALN), was calculated with a similar logic (see appendix 2).

INSERT TABLE 6 HERE

Finally, regarding *status of the Additional Language*, Welsh-English children growing up in Wales were coded as societal bilinguals, all others not.

***Measuring vocabulary*.** To measure children's vocabulary achievements in English and in their Additional Language for the 13 target languages, we used Communicative Developmental Inventories in each language. For the English CDI, we developed a 100-word version of the existing Oxford CDI (Hamilton et al., 2000), referred to as the Oxford Short Form CDI, by selecting words from the original 416 words which would (1) be representative of the words known and produced by 24-month-old monolinguals in the original norms that cover the same range of frequencies, and (2) contain the same distribution of syntactic categories (nouns, verbs, pronouns, etc). We selected 10 words understood and produced by 100% of 2-year-old monolingual toddlers as provided by the Oxford CDI database, then 10 words understood and produced by 90% of the same children, etc. Then we adjusted these words to include a proportion of nouns, verbs and function words similar to those found in the Oxford CDI (see appendix 5 for the full list). To verify the validity of this Oxford Short Form CDI, the parents of 134 monolingual children from the Plymouth area (including 72 girls) aged 10 to 26 months (mean age 17.9 months) completed both the short and the long CDI within a week (mean number of days between completions: 4.3 days, SD 5.5). Their mean score on the long Oxford CDI was 160.2 words in comprehension (out of 416; SD 119.7) and 80.3 in production (SD 107.9); their mean score on the Oxford Short Form CDI was 43.5 words in comprehension (SD 28.0) and 23.0 in production (SD 28.2). Children's scores in the two CDIs were highly correlated in comprehension ($r = .95$, $p < .0001$) and in production ($r = .86$, $p < .0001$).

We also compared the scores directly for the 100 words that were present on both the long and the 100-word versions of the Oxford CDI: monolingual children's parents reported higher scores on the Oxford Short Form CDI, both for comprehension (t (133) = 5.71, p < .0001, mean Oxford CDI score = 39.2%; mean Oxford Short Form CDI: 43.5%) and production (t (133) = 5.40, p < .0001, mean Oxford CDI score = 20.4%; mean Oxford Short Form CDI: 23.0%). This difference is likely due to a fatigue or attentional effect when having to fill in a CDI four times as long as the 100 word Oxford Short Form CDI. Across the two completions, parents reported the same outcome (known or unknown) for 85.8% of words in comprehension, and 92.6% in production. Correlations between the short and long CDIs for the 100 words were r = .95 and r = .98 for comprehension and production respectively (p < .0001), indicating excellent validity for the Oxford Short Form CDI.

For the Additional Languages, we used the adaptations of CDIs for 12 Additional Languages with the authors' permission (see list in the appendix 1), selecting the form adapted for the age of 24 months when multiple versions were available. Additional Language CDIs had lengths varying from 654 words in Greek (Kati, personal communication) to 62 in Bengali (Hamadani et al., 2010; see appendix 1 for references of CDIs).

We developed a new CDI for Hindi/Urdu as none were available. For simplicity we treated these two languages as dialects of the same language using different graphemic systems, so we developed the same version, written in the two alphabets. Following the method by Kern (2007), after a translation of the Oxford CDI, two focus groups of native Urdu speakers agreed on a cultural adaptation of the word list. Native Hindi speakers were consulted to check its adaptation to Hindi.

All parents were first asked to complete the Oxford Short Form CDI, assessing receptive and productive vocabulary separately (for each word, they were to assess whether the word was understood but not produced, or produced). If they felt unable to do so because, for example, they never spoke English at home and therefore could not estimate their child's English

knowledge, a proficient English speaking caregiver would complete a printed version of the CDI (e.g., a childminder). Parents were asked to complete the appropriate Additional Language CDI within a week of the completion of the Oxford Short Form CDI.