Journal of Child Language (2025), **52,** 1280–1294
doi:10.1017/S0305000924000473

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Frequency, redundancy, and context in bilingual acquisition

Paul Ibbotson[1], Stefan Hartmann[2], Nikolas Koch[3] and Antje Endesfelder Quick[4]

[1]School of Education, Childhood, Youth and Sport, The Open University, Milton Keynes, UK; [2]Faculty of Arts and Humanities, German Department, Heinrich-Heine-University Düsseldorf, Germany; [3]Institute for German as a Foreign Language, Ludwig-Maximilians-Universität München, Germany and [4]Department of British Studies, University of Leipzig, Germany
Corresponding author: Stefan Hartmann; Email: hartmast@hhu.de

(Received 22 January 2024; revised 18 April 2024; accepted 29 May 2024; first published online 12 December 2024)

Abstract

We report findings from a corpus-based investigation of three young children growing up in German-English bilingual environments (M = 3;0, Range = 2;3–3;11). Based on 2,146,179 single words and two-word combinations in naturalistic child speech (CS) and child-directed speech (CDS), we assessed the degree to which the frequency distribution of CDS predicted CS usage over time, and systematically identified CS that was over- or underrepresented in the corpus with respect to matched CDS baselines. Results showed that CDS explained 61% of the variance in CS single-word use and 19.3% of the variance in two-word combinations. Furthermore, the bilingual nature of the over or -underrepresented CS was partially attributable to factors beyond the corpus statistics, namely individual differences between children in their bilingual learning environment. In two out of the three children, overrepresented two-word combinations contained higher levels of syntactic slot redundancy than underrepresented CS. These results are discussed with respect to the role that redundancy plays in producing semiformulaic slot-and-frame patterns in CS.

Keywords: bilingualism; child-directed speech; redundancy; frequency

Introduction

A significant challenge in language acquisition research is explaining the variance between children in what they say, when they say it, and why. To that end, a large amount of evidence have shown that usage-based patterns in child-directed speech (CDS) explain a significant proportion of the variance in child speech (CS). For example, differences in the way caregivers structure their communicative interactions have been shown to predict differences in children's lexical frequency, speech errors, age-of-acquisition effects, and multiword chunk use (Ambridge et al., 2015; Diessel, 2007; Lieven et al., 1997; Matthews & Bannard, 2010; Odijk & Gillis, 2021; Quick et al., 2021).

Because language is a conventionalised agreement between speakers on how language should be used, anyone who learns a language is by definition also learning its patterns of

use. So, all else being equal, we would expect CS usage patterns to gradually approximate those in CDS as children grow older. On its own, this fact tells us little about *how* children are reaching this approximation. In principle, any number of different learning mechanisms could drive two statistical distributions to overlap throughout development. What is arguably of more interest therefore are cases that deviate from this expectation. For instance, the following factors have all been shown to explain some CS variance not explainable by CDS variance: how easy a word is to segment from the speech stream and articulate (Christophe & Dupoux, 1996; Monaghan & Christiansen, 2010; Vihman & Vihman, 2011); a word's imageability, semantic transparency or grammatical class (Bird et al., 2001; Gentner, 1982; Narasimhan & Gullberg, 2011); whether there is an easily identifiable referent (Gentner, 1982); whether the word occurs during episodes of joint attention (Tomasello & Farrar, 1986); is directed to the child rather than overheard in speech (Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013); or how knowledgeable the speaker is (Sabbagh & Baldwin, 2001). In these cases, deviations from the expected CDS-CS relationship point to how children are using different social and cognitive sources of information to construct their language.

To provide a systematic way for the language acquisition researchers to identify these deviations, Ibbotson et al. (2018) developed an open-source analytic tool: the frequency filter. This tool assumes a strong CDS-CS relationship as its starting point and then identifies outliers that are statistically different from this expectation. When words and phrases are over- or underrepresented in the child's speech relative to the language they hear, it indicates a process at work that is above and beyond that of frequency and so may be a candidate for a cognitive or social developmental process. Over- or underrepresented language use is formally defined by the size of the residual value following a CDS-CS regression (more detail in Methods). Ibbotson and colleagues applied the frequency filter to naturalistic corpora of English, Swedish, Japanese, French, Italian, and Spanish and found that first-person language use, the learnability of nouns versus verbs, and a noun's imageability significantly shaped CS acquisition trajectories above and beyond CDS frequency (and above and beyond between language differences too; 2018). For example, young children had a bias to use the first-person singular language of *I*, *me*, *my*, *mine*, before they learned to use the plural discourse of *we*, *us*, *our*, which takes a more collaborative orientation. Importantly, these effects were established *after* controlling for CDS input, and so the differences in CS use is more likely to reflect the developing perspective-taking skills and communicative motivation of the child (a finding recently replicated by Vasil et al., 2023). The fact that the same pattern emerged across different languages, with different morphosyntactic and lexical resources, provides stronger evidence for a common social-cognitive bias.

The frequency filter methodology offered a promising proof of concept but was limited in its scope: it was performed on single words, on particular word classes, and with monolingual children. In the current study, we adapt and extend the frequency filter methodology in four novel ways. (1) we apply it to the entire corpus of speech, whereas before it has only been applied to various word classes such as pronouns, verbs, and nouns. (2) We apply it to two-word combinations, whereas before it has only been applied to single words, allowing us to examine multiword chunks for the first time using this method. (3) We apply the frequency filter to children growing up in bilingual environments, whereas before it has only been applied in monolingual contexts, and (4) we include a redundancy analysis of over- and underrepresented speech.

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

Our redundancy analysis is motivated by the fact that in usage-based linguistics, the repetition or redundancy of items in lexical frames plays an important role in the schematisation of more abstract syntax. Put simply, the most frequently co-occurring items in speech are candidate sites for semiformulaic slot-and-frame patterns that pave the way for more complex subject–predicate constructions that emerge later in development (Theakston et al., 2015; Tomasello, 2003). These schemas are likely to emerge where a number of different lexical items (*X*) occur in a variable slot alongside a relatively frequent and redundant element (the frame), for example, "Where's the *X*?," "I wanna *X*," "More *X*," "It's a *X*," "I'm *X*-ing it," "Put *X* here," "Mommy's *X*-ing it," "Let's *X* it," "Throw *X*," "*X* gone," "I *X*-ed it," "Sit on the *X*," "Open *X*," "*X* here," "There's a *X*," "*X* broken." To capture this usage pattern, we calculate the redundancy of two-word combinations in CS using a measure of how many times a lexical item is repeated in a given slot (more detail in Methods). As schematisation requires repeated experience with slot-and-frame variation, we would expect redundancy to be relatively higher in CS that is overrepresented compared to that which is underrepresented, as defined by the frequency filter. That is, CS language that is frequent enough to be identified as an outlier from CDS is more likely to contain productive slot-and-frame patterns.

As useful as the frequency filter is, and frequentist approaches are more generally, it is only as good as the corpus statistics. There are of course a broad range of environmental factors that bear on language acquisition that are not necessarily captured by the corpus itself. These include individual differences between family structure, such as the presence of siblings and grandparents and the changing circumstances that affect children over time, such as periods spent abroad, time spent in kindergarten, and with friends. The role of environmental context is especially relevant in bilingual acquisition where it is notoriously difficult to find two bilingual children exposed to exactly the same L1/L2 ratio for a significant proportion of time. In this study, we make use of this broader environmental context to explain some of the usage-based patterns in bilingual acquisition.

To summarise, based on previous work on the nature of CDS-CS (e.g., Diessel, 2007; Tomasello, 2003), and prior work with the frequency filter method specifically (Ibbotson et al., 2018), we predict that (1) there will be a significant association between CDS and CS frequency distributions and the strength of this relationship will increase over development for both one-word and two-word combinations; (2) because the frequency filter controls for CDS but is limited to the corpus data, broader environmental context will be important in explaining individual differences in the bilingual nature of CS; and (3) based on role of frequency in schematisation, overrepresented CS will contain more productive two-word combinations compared with underrepresented CS.

Methods

Participants

Our data consist of the naturalistic speech of three children, Fion, Silvie, and Lily ($M = 3;0$, range = 2;3–3;11), and their parents. Recordings took place at home while the children carried out their usual day-to-day activities of playing, mealtimes, reading books, and bedtime. All three children grew up as German-English simultaneous bilinguals from birth, with one parent who was an L1 speaker of German and the other an L1 speaker of English and all were living in a German city. Beyond these similarities, there were significant differences in both the relative dominance of L1/L2 and the broader environmental context in which they heard CDS.

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

Fion

Fion was recorded between 2;3 and 3;11 for an average of 2 hours per week. His mother is an L1 speaker of German and his father an L1 speaker of English. Both parents spoke the L2 language quite well, but they were very inconsistent in their use of their mother tongues, often addressing each other using their L1 and L2 languages interchangeably. Fion had an older brother and the parents and brother sometimes produced code-mixed utterances. From early on, Fion was exposed to German at home from his mother for most of the day, and then later on at a German-speaking day-care centre, for 4 hours a day at 19-months-old and then 6–8 hours a day from 24 months-old onwards. Until the age of three years, Fion's major input was German, but after his third birthday, there was a shift towards more English with an extensive stay in the father's home country and more frequent visits by his English-speaking grandparents who did not speak or understand any German.

Silvie

Silvie was recorded between 2;4 and 3;9 for an average of 2.5 hours per week. Her mother is an L1 speaker of English and her father is an L1 speaker of German. Silvie's mother was her only source of English and when the mother addressed Silvie directly she used English. Silvie also heard her mother speak German when the father was around or when they were outside their home environment, and thus, she knew that the mother understood and spoke German perfectly well. Additionally, the parents also converged on German as the family language since the father's proficiency in English was rather limited. Silvie's mother stayed home with her child during the first, and most of the second year and during this time, Silvie was mainly exposed to English during the day. However, from 18 months onwards, Silvie's input situation changed as she started attending a German kindergarten for 45 hours per week and was mostly exposed to German from that time on.

Lily

Lily was recorded between 2;3 and 3;10 for an average of 1 hour per week. Just as in Silvie's case, the mother was the English L1 speaker and the father was a German L1 speaker. However, since both parents spoke each other's language very well, they did not settle on a family language but instead used both German and English interchangeably. Lily's brother was also raised as a simultaneous bilingual and so he provided further input in German, English, and occasionally in code-mixed speech. Lily stayed home during the first year with her English-speaking mother and entered a German kindergarten at the age of 18 months for most of the day. Although Lily's input situation resembles Silvie's, her home language distribution was much more balanced.

To summarise the language context, all three children were raised in Germany as German-English bilinguals, and all families subscribed to the idea of one parent one language household, but how they implemented the idea in practice varied a lot. Fion's English input came mostly from his father, but he also frequently heard his mother and brother speaking English. Both parents were very inconsistent in their use of their L1 language. After his third birthday, Fion's input to English increased thanks to frequent visits from his English-speaking grandparents and a long vacation in his father's home country. Silvie had only one source of English but various sources of German, and German was the family language. Lily, on the other hand, had various sources of English

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

and German, and experienced both her parents interacting in their L2 as well. Both languages were used as the family language.

Transcription and data annotation

All recordings were transcribed and coded in the SONIC CHAT format (MacWhinney 2000) by a bilingual research assistant. Nonstandard variants and speech errors were tagged with annotations indicating the standard variant, e.g., English *gonna* [: *going to*], German *bei den* [: *dem*] *Traktor* "at the tractor". This allows us to work with the original utterances (*gonna, bei den Traktor*) or with the normalised ones (*going to, bei dem Traktor*). For the present study, only the normalised data were taken into account, as we are interested in the alignment between children's and adult's use of lexical items. Table 1 gives an overview over the size of the dataset.

Procedure

Altogether, the data from the above corpora provided 2,146,179 tokens of naturalistic CS and CDS, which we analysed according to the following procedure.

The frequency filter methodology begins by counting the frequency of items that occur in both CS and CDS for a given month in the corpus. For example, the word "me" might be said 15 times by the child and it might be spoken 25 times in CDS at the child's age of 2;5. Note that because we are calculating frequency of use rather than simply if a word occurs at all, as may be the case for studies of age-of-acquisition (e.g., Roy et al., 2015), the only forms included are words or two-word combinations that occur in both child and caregiver speech at least once. This has the advantage of ensuring that any significant effects are not attributable to vocabulary differences between children and their caregivers. On the other hand, this restriction is likely to exclude more speech errors, simply because language-learning children tend to make more speech errors than their parents. To be clear, the frequency filter does not directly filter errors out, but having the stipulation that an item needs to be attested by *both* parent and child, simply means that they are less likely to be included in the analyses. The methodological trade-off between gaining vocabulary control but losing speech errors is a point to which we return in the limitations section of the Discussion. The single-word and two-word frequency lists that we used were obtained from the transcribed corpora, drawing on the normalised utterances as described above. From each dataset, single words and two-word

Table 1. Overview of the number of utterances in the three datasets. The CS data have been tagged for whether they are German, English, or code-mixed. Differences to 100% were categorised as ambiguous (e.g., one-word utterances like *hm?* that cannot be clearly assigned to one of the two languages.)

		Fion	Silvie	Lily
CDS		180,293	139,993	146,909
Child speech	Sum	47,812	37,995	60,184
	German	34,837 (72.9%)	28,294 (74.5%)	10,597 (40.3%)
	English	9,467 (19.8%)	5,415 (14.3%)	10,123 (38.5%)
	Code-mixed	3,508 (7.34%)	4,286 (11.2%)	2,218 (3.43%)

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

combinations were extracted using an algorithm that crawls through the corpus utterance by utterance. For instance, a sentence like *the cat is hiding under the mat* would be segmented into single words as follows: |*the*|*cat*|*is*|*hiding*|*under*|*the*|*hat*. For two-word combinations, the same utterance would be parsed as |*the cat*|*cat is*|*is hiding*|*hiding under*|*under the*|*the mat*. Unintelligible speech (transcribed as *xxx*) was not taken into account.

Because of the natural Zipfian distribution of the language data (Zipf, 1935/1965), once a frequency list has been generated for a particular time point, a log transform (ln (Language)/ln(10)) is performed on both CS and CDS separately, which rescales the values: 1 becomes 0, 10 becomes 1, 100 becomes 2, and so on, and produces improved heterogeneity of variance, preparing the data for the regression analysis to come. This procedure was the same for single words and two-word combinations.

Following the log-transformation, the frequency counts are entered into a simple linear regression with CDS as a continuous predictor variable and CS as an outcome variable. The regression generates two statistics of interest here. First, it provides an overall estimate of the strength of the relationship between CDS and CS (expressed as R^2 in the Results section). Note that by using relative frequency rather than absolute frequency, we control for the rather prosaic fact that caregivers use more language overall (both types and tokens) than children.[1] Using a correlational approach allows us to focus on the relationship between the CDS and CS in the rank order token frequency of types. For example, in the situation where a child says "me" much more frequently than we would expect given how much they have heard "me", and given the relationship of all other items in the corpus. When the regression analysis is repeated for each month of the corpus, this R^2 value gives us a dynamic measure of how the CS frequency distribution is coming to approximate CDS frequency distribution over the course of development. Recall this kind of approximation over time is what we would expect, given that anyone who learns a language is by definition also learning its patterns of use. This brings us to our second statistic of interest: residuals.

All other things being equal, the regression line is a good predictor of CS use across a number of parts of speech and languages (e.g., Diessel, 2007; Tomasello, 2003; Ibbotson et al., 2018). As has been demonstrated elsewhere, however, all other things are not equal; there are a number of social and cognitive factors that can pull individual lexical items away from the regression line, for example, some words are more salient in the speech stream than others. So, as well as an R^2 value, the regression analysis also generates standardised residual values (ZResiduals) for each item entered into the analysis. The larger these residual values, the further the item is from the regression line. Therefore, items with the largest residual values are the outliers from the expectation that CDS is a good predictor of CS. Negative ZResidual values indicated CS<CDS or to put it another way, CS is underrepresented, and positive ZResidual values indicated CS>CDS or CS is overrepresented. When these residuals are ranked from smallest to largest, we have a systematic way of easily identifying the outliers.

In this study, we chose as the top 25 and bottom 25 outliers as our cut-off point. By choosing the most extreme cases first and working backwards toward regression line, we gave ourselves the best chance of detecting any differences that would be of interest. When

[1]We are aware of the fact that this approach still entails some simplifications. Perhaps most importantly, given the properties of word frequency distributions, the relative frequency of the same word can differ considerably depending on the sample size (Baayen, 2001), and quite naturally, the CS data are much more sparse than the CDS data. But in light of the fact that we are mainly interested in the input-output relationship, we argue that despite these simplifications, our approach can give valuable insights.

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

these 50 items per month were multiplied by the duration of the corpus, it gave us good coverage of the extremes, yet still manageable given the labour-intensive manual input of the methodology (50 words*21 months for Fion; 50 words*18 months for Silvie; and 50 words *19 months for Lily).

The output of the previous steps in the procedure gives over- and underrepresented CS language for each child and for each month's worth of data in the corpus. Next, we explored the linguistic nature of these outliers in two ways. First, we examined whether over- or underrepresented usage patterns can be characterised as either German, English, or a mixture of two, and whether these patterns change over time. So, for each item, for each month's worth of corpus data, we labelled whether the outlier was either German or English. All children were raised as German-English bilinguals, but as noted, with significant differences in both the relative dominance of L1/L2 and the broader environmental context in which they heard CDS. Because the parent input is matched to the child, we can be sure any corpus-based parent variability is taken into account by the frequency filter methodology. That is, if one parent says a word particularly frequently and so does the child, this will likely not show in the residual outliers. Thus, any dominance for either German or English in children's over- or underrepresented speech is established *after* controlling for their bilingual input environment in the corpus.

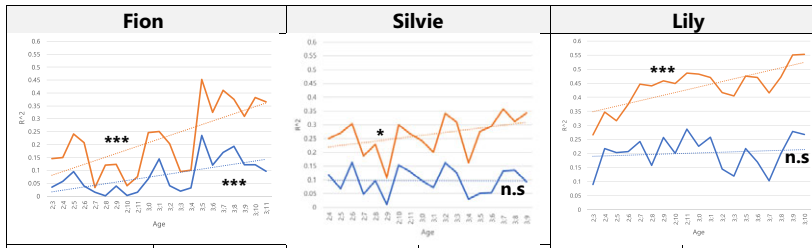
Second, we explored whether two-word combinations are more or less redundant depending on whether they are either over- or underrepresented. Redundancy is a measure of the repetition or predictability of a source of information. We calculate redundancy as the total overlap of lexical items in a given slot. For example, if the child says "ich hab," "ich mache," "ich nehme," "ich nicht," and "du bist," then the first slot has 3 repetitions of the same word but the second slot has zero. The basic idea is that the redundancy in the first slot is a candidate site around which a frame will form and so over time, the child comes to represent the above sequence as "ich_*X*" and "du-bist" (a full example from Fion's data is given in the Appendix with further detail about how this measure is calculated). For this reason, the redundancy analysis focuses on the two-word combinations extracted from our corpus rather than the single words. We performed the redundancy calculation for both slots of the two-word combinations, which gave us an average redundancy measure for 25 most over- and underrepresented examples of CS. As before, we repeat this process for each month's worth of data, and it gives us a dynamic measure of how redundancy changes over time.

Results

We present the results in the order we introduced in the procedure. First, we show the overall strength of the relationship between CDS and CS and then we examine the nature of the outliers from this relationship, both in terms of German-English characteristics and then in terms of redundancy.

Figure 1 shows that on average CDS explains 61.3% of the variance in CS single-word use and 19.3% of the variance in two-word combinations. The values on the y-axis show the strength of relationship between CDS and CS for one-time point, and when we lay these values end-to-end, for the entirety of the corpus, we can see whether there is any trend in this relationship over time. Thus, the p-values represent a kind of meta-correlation; the correlation of strength of CDS-CS relationship over time. For Fion, Silvie, and Lily the strength of the relationship between CDS and CS single-word frequency distributions significantly increases over time ($p<.001$, $p<.05$, $p<.001$). For two-word

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press



		CS	CDS	CS	CDS	CS	CDS
Single words		117,738	396,954	115,964	490,563	137,976	443,373
		(N=21) R^2 = .67, p<.001		(N=18) R^2 = .42, p<.05		(N =19) R^2 = .75, p<.001	
Two words		34,700	63,110	37,504	74,049	60,299	113,650
		(N=21) R^2 = .58, p<.001		(N=18) R^2 = -.01, p=.96		(N =19) R^2 = .12, p=.59	

Figure 1. Top graphs represent the proportion of CDS frequency variance that explains CS frequency variance over the course of development, single words = orange line, two-word combinations = blue line, linear trends = dotted lines. Numbers in the table represent the total number of tokens entered into the analyses and the R^2 expresses the strength of the association between CDS and CS, and whether it significantly changes over time (*p*-value).

combinations, this result also holds for Fion (p<.001) but not for Silvie and Lily (p=.96, p=.59).

For Fion and Silvie, it appeared that their overrepresented speech was dominated by German, and their underrepresented speech was dominated by English, for both single and two-word combinations. For Lily it seemed that her speech was much more balanced between German and English. To assess the statistical significance of these patterns, we performed a two-tailed Chi-square because over/underrepresented, German/English are defined here as binary mutually exclusive categories. The output of this analysis shows us whether there are significant dissociations between language and outlier type (Table 2).

The results show a significant dissociation for Fion and Silvie such that underrepresented CS is much more likely to be English than German, and overrepresented CS is much more likely to be German than English; this pattern holds for single words and two-word combinations. There are also significant dissociations for Lily, but these are in the opposite direction. Her underrepresented speech is much more likely to be German than English, and overrepresented speech is much more likely to be English than German; this pattern holds for single words and two-word combinations.

Redundancy of outliers

Finally, we turn to the redundancy of the two-word combinations, broken down by under- and overrepresented speech (Figure 3).

For reasons outlined in the Introduction, a key theoretical question is whether overrepresented CS is more redundant than underrepresented speech. To assess the statistical difference of the redundancy distribution, we conducted an independent t-test between over- and underrepresented speech. For both Fion and Silvie, overrepresented language was significantly more redundant than underrepresented language (p<.001, p<.001). Or

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

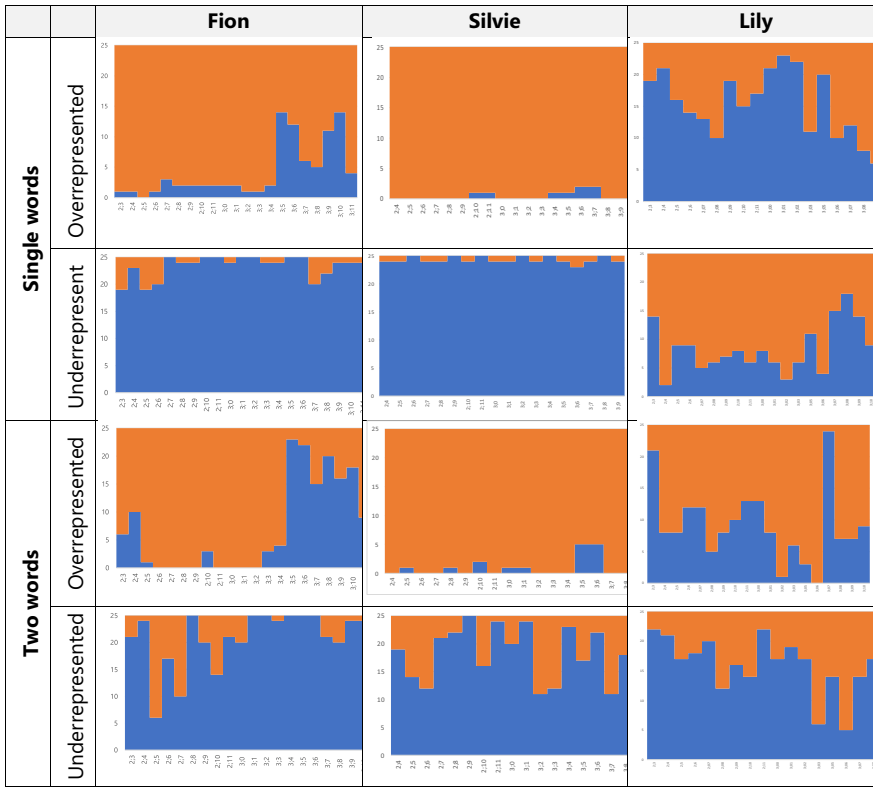


Figure 2. Bilingual character of the outliers over the course of development. German = orange, English = blue. *y*-axis =outliers, and *x*-axis =age.

Table 2. The disassociation of language and under- and overrepresentativeness by single words and two-word combinations and associated Chi-squared statistics.

		Fion		Silvie		Lily	
		English	German	English	German	English	German
Single words	Overrepresented	88	437	8	442	284	191
	Underrepresented	490	35	437	13	160	315
		χ^2 (1)=621.973, p<.0001		χ^2 (1)=818.061, p<.0001		χ^2 (1)=65.018, p<.0001	
Two-words	Overrepresented	150	375	24	426	290	185
	Underrepresented	434	91	336	114	175	300
		χ^2 (1)=311.91, p<.0001		χ^2 (1)= 450.667, p<.0001		χ^2 (1)= 55.70, p<.0001	

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

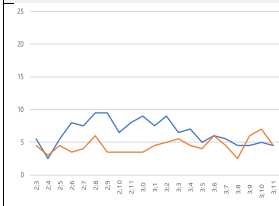
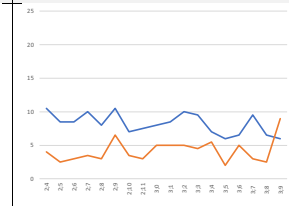
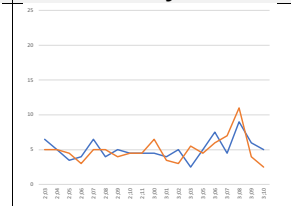
Fion	Silvie	Lily
		
t(40) = 4.21, p<0.001; Mean overrepresented redundancy = 6.5 (SD =1.91); Mean underrepresented redundancy = 4.45 (SD = 1.13) UCI: 3.02; LCI: 1.06, d = 1.57.	t(34) = 7.44, p<0.001; Mean overrepresented redundancy = 8.22 (SD =1.52); Mean underrepresented redundancy = 4.19 (SD = 1.71). UCI: 5.12; LCI: 2.92, d = 1.62	t(36) = 0.4, p=0.71; Mean overrepresented redundancy = 5.07 (SD =1.49); Mean underrepresented redundancy = 4.94 (SD = 1.87). UCI: 1.4; LCI: -.98.

Figure 3. Redundancy of CS two-word combinations over the course of development. Overrepresented use = blue line and underrepresented = orange line. As before, the y-axis is limited to 25 because we limited ourselves to the 25 highest and lowest residuals.

in other words, overuse of forms in CS was associated with more frame-and-slot like productivity in two-word combinations. For Lily, there was no evidence that either over- or underrepresented language significantly differed in terms of redundancy.

Discussion

We explored the relationship between CDS and CS in the context of three children acquiring German and English in bilingual households. Beyond these similarities, there were significant differences in both the relative dominance of L1/L2 and the broader environmental context in which the children heard CDS. After applying the frequency filter to a corpus of speech, we explored the nature of CDS-CS outliers to determine their bilingual character and the productivity of two-word combinations as assessed through a redundancy measure. In what follows, we return to each of our predictions made in the Introduction.

(1) There will be a significant association between CDS frequency distribution and CS frequency distributions, and this increases over development for both one-word and two-word combinations. We found that averaging across all children, for all developmental periods, CDS explains 61.3% of the variance in CS single-word use and 19.3% of the variance in two-word combinations. For Fion, Silvie, and Lily, the strength of the relationship between CDS and CS single-word frequency distributions significantly increased over time. For two-word combinations, this result also holds for Fion, but not for Silvie and Lily. As laid out in the Introduction, there is a sense in which this is to be

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

expected. By definition, part of learning a language involves a shared understanding of how language is used, so all else being equal, the usage patterns of CS should come to approximate those in CDS as children grow older. Before moving on to what we consider the more theoretically interesting case of the outliers to this relationship, it is worth considering two points.

First, the CDS-CS relationship is demonstrated here in the context of the entire corpus of speech, whereas before it has only been established in the context of pronouns, verbs, or nouns by Ibbotson and colleagues (2018). In that study, looking at English monolingual pronoun use, CDS explained about 55% of the variance in CS use. The fact that we see similar levels of variance explained in our bilingual corpora when all parts of speech are considered speaks to the robustness of this relationship across different contexts, including bilingual acquisition. Second, it is noticeable that the two-word R^2 values track that of single words, but at a consistently lower level (Figure 1). This was not anticipated but is perhaps not too surprising for the following reasons. The longer the string of words, the rarer that string of words will be in a corpus. For example, we are likely to find more tokens of "dog" than "the dog under the table is reading a newspaper," thus as an utterance gets longer, the compositionality of language reflects a more niche communicative intention. In our corpus analysis, CS single words have a higher token frequency than two-word combinations by about a ratio of 3:1. So, children have more opportunities in general to approximate CDS frequency of single words compared with two-word combinations. This fact might help to explain why for all children, two-word combinations have an R^2 that is consistently lower than single words.

We turn now to the exploration of the outliers and our second prediction (2) because the frequency filter controls for CDS but is limited to the corpus data, broader environmental context will be important in explaining individual differences in the bilingual nature of CS. Figure 2 and Table 2 show a significant dissociation such that underrepresented language is much more likely to be English than German; overrepresented language is much more likely to be German than English; and this pattern holds for single words, and two-word combinations for Fion and Silvie. Lily on the other hand shows dissociation in the opposite direction. Her underrepresented language is much more likely to be German than English; overrepresented language is much more likely to be English than German; and this pattern holds for single words and two-word combinations.

Because the parent input is matched to the child, we can be sure that any between-parent corpus-based variability is taken into account by the frequency filter methodology. For this reason, to explain individual differences, we need to turn to effects above and beyond the CDS frequency and look at the broader environmental context. In this regard, it is noticeable that Fion's approximation to CDS increases rapidly at 3;5 (Figure 1), and his overrepresented language use also becomes more balanced between English and German (Figure 2) at the same point in development. Until the age of three years, Fion's major input was German, but this changed after his third birthday with a shift towards more English with an extensive stay in his father's home country and more frequent visits by his English-speaking grandparents who did not speak or understood any German. This boost in English for Fion (but not Silvie or Lily) corresponds with the timing of Fion's approximation to more English use around 3;5. For Fion and Silvie, despite being raised in a bilingual home, the evidence suggests that they are much more likely to go beyond the input baseline with German rather than English. This is likely due to the fact that their bilingual environment at home was not offset by the German dominance heard outside of the home. The shift in language use at 3;5 would be extremely difficult to explain by looking at the corpus statistics in isolation and needed an understanding of the broader contextual factors affecting Fion's development.

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

Lily's dissociation is in the oppositive direction to that of Fion and Silvie and, to explain such a difference, it is also helpful to consider a broader environmental context than the corpus statistics. Both of Lily's parents spoke each other's language very well, and they did not settle on a family language but instead used both German and English interchangeably. Lily's brother was also raised as a simultaneous bilingual and so he provided further input in German, English and occasionally in code-mixed speech. Although Lily's input situation resembles Silvie's, her home language distribution was much more balanced. This is reflected in much more balanced German-English values of under- and over-represented speech (as indicated by Lily's lower Chi-squared values in comparison to Fion and Silvie, Table 2). It is also noticeable that Lily shows the strongest association between CDS and CS (Figure 1) with up to 55% of the variance in her CS explainable by CDS. This shows that Lily most closely approximates the frequency distribution of speech she hears around her, and that speech is the most balanced between German-English of the children we consider here.

Finally (3) based on the role of frequency in schematisation, overrepresented CS will contain more productive two-word combinations compared with underrepresented CS as in [da ist *X*] [there is *X*] or [ich bin *X*] [I am *X*]. This prediction was of theoretical interest because in usage-based linguistics, the most frequently co-occurring items in speech are candidate sites for semiformulaic, productive schemas that pave the way into more complex and abstract subject–predicate constructions that emerge later in development (Theakston et al., 2015; Tomasello, 2003). We found that for both Fion and Silvie, overrepresented language was significantly more redundant than underrepresented language. Or in other words, after controlling for CDS input, overuse of forms in CS was associated with more productivity in two-word combinations. At least for Fion and Silvie, we established that overrepresented language is best characterised as German, and underrepresented as English. Therefore, we can also conclude that, in general, German is used more productively than English, for both Silvie and Fion.

As before, Lily showed a different pattern to the other children where there were no significant differences in the redundancy of over- versus underrepresented language. We can only speculate as to why, but one plausible possibility is that the more balanced German-English input Lily receives, means that she is slower to reach productivity in either German or English. That is, an hour spent accumulating the statistical redundancies of two-word patterns in one language is an hour not spent learning them in the other language. With only so many hours in the day, this protracts the developmental trajectory towards productivity in both languages, although the overall productivity levels may be no different than monolinguals later on in development. Had the corpus extended beyond 3;09 for Lily, this is a possibility we could have assessed.

At this point, we need to introduce a point of caution about equivalating redundancy with productivity. We have argued that two-word combinations with high levels of redundancy, especially in overrepresented speech, are candidate sites where productive language use is likely to emerge in development. However, because we base our analysis entirely on naturalistic speech, we cannot rule out that what looks like productivity is actually the result of using highly entrenched multiword chunks, rather than flexibly using slot-and-frame type patterns. To pull apart these possibilities, we would need to demonstrate flexibility beyond the input, which is something we are not able to do given the restriction that two-word combinations need to have been attested in both CDS and CS. The gold-standard test here would be to elicit productive use in a two-word combination with a nonce word, which has a frequency of zero. If children were able to do that, it suggests some productive schema where novel items can be dropped into a

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

frame if they satisfy the right requirements. In this regard, the frequency filter could be very useful in generating a naturalistic list of potential productive schemas that could be confirmed or refuted by later experimentation with nonce words; the prediction being that children should much more readily substitute a nonce word with overrepresented speech than underrepresented speech.

Another limitation that the frequency filter approach suffers is that bigrams may not always be a reliable way of investigating the relationship between input and output as early child language is characterised by phenomena such as ellipsis of function words (e.g. *want sandwich* instead of *I want a sandwich*).[2] Future refinements of the frequency filter approach could circumvent this problem by working with a broader set of n-grams including skip-grams, i.e., word combinations that are not directly adjacent but appear within a specific window (e.g., __ *want a sandwich*, *i* __ *a sandwich*, *i want* __ *sandwich*, and *i want a* __ would be skip-grams of *i want a sandwich* with 3-grams in a window of 4 words).

Any conclusions we draw here are also caveated by the fact that we have only been able to analyse three children and their parents, albeit using 2,146,179 tokens of naturalistic CS and CDS. We have also limited the analysis to mainly conventional speech, because each item needs to have been uttered at least once in CS and CDS to enter into the analysis. This has the advantage of ensuring that any significant differences we find are not attributable to differences in what words children and caregivers know because an item needs to have occurred at least once in both CDS and CS. But we also acknowledge that this requirement imposes the disadvantage of being more likely to exclude speech errors from the analysiss because instances of unconventional use in CS are less likely to be attested by at least one example in CDS. Another caveat is that the language shift that Fion experienced in his input (see Section "Participants" above) is not reflected in the recorded data: his stay with the English-speaking grandparents corresponds to a small gap in the recorded data, which is why the language proportions in the recorded input data remain fairly stable. This is, however, a general problem of corpus-based methods that can only ever sample a small and not always fully representative proportion of an individual's linguistic input.

In summary, we found (1) strong support that the single-word frequency distribution of CS approximates that of CDS in a bilingual corpora and partial support that it does so for two-word combinations (2) that the broader environmental context can be used explain some of the bilingual usage patterns above and beyond the statistics of the corpus and (3) support in two out of the three children studied that overrepresented CS is more likely to contain slot-and-frame patterns than underrepresented CS, underlining the theoretical importance of the relationship between frequency and redundancy in language development.

Data availability statement. The full corpora cannot be made available yet as they contain sensitive personal data. However, Antje Endesfelder Quick, Nikolas Koch and Stefan Hartmann are currently working on pseudonymising the data and making them available via the CHILDES database. The Frequency Filter script is publicly available, see Ibbotson et al. (2018) for details.

Acknowledgements. Thank you to the parents and children who took part in this study and to the anonymous reviewers who significantly improved the manuscript.
The research reported here was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 504095269. In addition, our research was conducted in the context of the scientific network "Language contact phenomena in multilingual language acquisition" (LaCoLA), also funded by the DFG, project number 496468900.

[2]Thanks to an anonymous reviewer for drawing our attention to this.

Competing interest. The authors have no competing interests to declare.

References

Ambridge, B., **Kidd, E.**, **Rowland, C.**, & **Theakston, A.** (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, **42**(2), 239–273.

Baayen, R. H. 2001. *Word frequency distributions (Text, speech and language technology)*. Dordrecht: Kluwer.

Bird, H., **Franklin, S.**, & **Howard, D.** (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments& Computers*, **33**, 73–79.

Christophe, A., & **Dupoux, E.** (1996). Bootstrapping lexical acquisition: The role of prosodic structure. *Linguistic Review*, **13**, 383–412.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, **25**(2), 108–127

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development, vol 2: Language, thought, and culture* (pp. 301–334). Hillsdale, NJ: Erlbaum.

Ibbotson, P., **Hartman, R.** & **Björkenstam, K.** (2018). Frequency filter: An open access tool for analysing language development. *Language, Cognition and Neuroscience*, **33**(10), 1325–1339.

Lieven, E. V. M., **Pine, J. M.**, & **Baldwin, G.** (1997). Lexically based learning and early grammatical development. *Journal of Child Language*, **24**, 187–219.

MacWhinney, Brian. (2000). *The CHILDES project: Tools for analyzing talk*. 3rd edn. Hillsdale and N.J: Erlbaum.

Matthews, D. & **Bannard, C.** (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science*, **34**: 465–488.

Monaghan, P., & **Christiansen, M. H.** (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, **37**, 545–564.

Narasimhan, B., & **Gullberg, M.** (2011). The role of input frequency and semantic transparency in the acquisition of verb meaning: Evidence from placement verbs in Tamil and Dutch. *Journal of Child Language*, **38**, 504–532.

Odijk, L., & **Gillis, S.** (2021). Fine lexical tuning in infant directed speech to typically developing children. *Journal of Child Language*, **48**(3), 591–604.

Quick, A. Backus, A. & **Lieven, E.** (2021). Entrenchment effects in code-mixing: Individual differences in German-English bilingual children. *Cognitive Linguistics*, **32**(2). 319–348.

Roy, B. C., **Frank, M. C.**, **DeCamp, P.**, **Miller, M.**, & **Roy, D.** (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(41), 12663–12668.

Sabbagh, M. A., & **Baldwin, D. A.** (2001). Learning words from knowledgeable versus ignorant speakers: Links between pre-schoolers' theory of mind and semantic development. *Child Development*, **72**(4), 1054–1070.

Shneidman, L. A., & **Goldin-Meadow, S.** (2012). Language inputand acquisition in a Mayan village: How important is directed speech? *Developmental Science*, **15**(5), 659–673.

Theakston, A., **Ibbotson, P.**, **Freudenthal, D.**, **Lieven, E.** & **Tomasello, M.** (2015). Productivity of noun slots in verb frames, *Cognitive Science*, **39**(6), 1369–1395.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M., & **Farrar, M. J.** (1986). Joint attention and early language. *Child Development*, **57**, 1454–1463.

Vasil, J., **Moore, C.**, & **Tomasello, M.** (2023). Thought and language: Association of groupmindedness with young English-speaking children's production of pronouns. *First Language*, **43**(5), 516–538.

Vihman, M. M., & **Vihman, V. A.** (2011). From first words to segments: A case study in phonological development. In I. Arnon & E. V. Clark (Eds.), *Experience, variation, and generalization: Learning a first language* (pp. 109–133). Amsterdam: John Benjamins

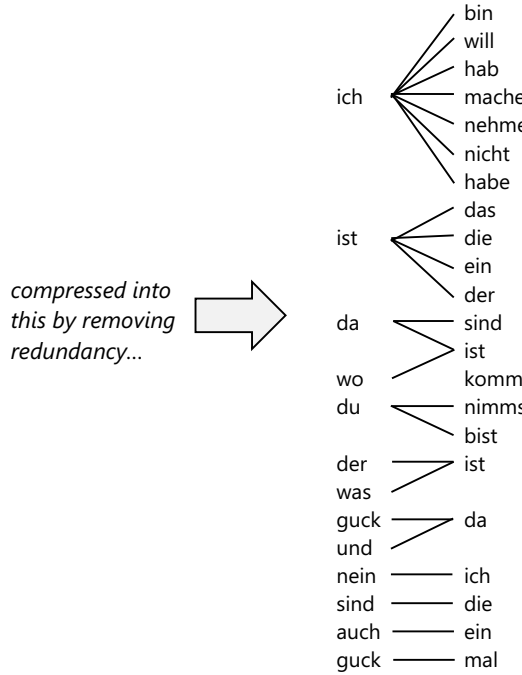
Weisleder, A., & **Fernald, A.** (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, **24**(11), 2143–2152.

Zipf, G. K. (1935/1965). *Psycho-biology of languages*. Cambridge, MA: MIT Press.

https://doi.org/10.1017/S0305000924000473 Published online by Cambridge University Press

Appendix 1

da	–	ist
ist	–	das
was	–	ist
ist	–	die
ich	–	bin
da	–	sind
ist	–	ein
guck	–	da
und	–	da
ich	–	will
nein	–	ich
der	–	ist
ich	–	habe
ist	–	der
du	–	nimmst
guck	–	mal
sind	–	die
wo	–	kommt
auch	–	ein
wo	–	ist
ich	–	hab
ich	–	mache
ich	–	nehme
ich	–	nicht
du	–	bist



Columns on the left represent two-word combinations extracted from the corpus of Fion's speech at 2;8; columns on the right are the same two-word combinations after the lexical redundancy has been removed and the data have been compressed. In the first slot, the number of lexical items has been reduced from 25 to 13, by compressing those 12 items which are redundant, for example, compressing 7 instances of "ich" into 1 "ich." In the second slot, we repeated the same procedure, which in this case reduced the items from 25 to 22. For each time period, we then took an average of the two slots, so for this example it was calculated as (13 +22)/2 = 17.5 out of 25. This procedure was applied separately for under- and overrepresented speech, and then repeated for each time slot and for each child.

Cite this article: Ibbotson, P., Hartmann, S., Koch, N., & Endesfelder Quick, A. (2025). Frequency, redundancy, and context in bilingual acquisition. *Journal of Child Language* **52,** 1280–1294, https://doi.org/10.1017/S0305000924000473