Assessing the impact of frequency and local adaptation mechanisms on child-caregiver language: a recurrence-quantificational approach

Robert Grimm
Universiteit Utrecht
Utrecht Institute of Linguistics OTS

M.A. Thesis

June 2014

First Supervisor: Dr. Raquel Fernández
Second Supervisor: Dr. Rick Nouwen
Third Reader: Dr. Luisa Meroni

Contents

1 Introduction

An important question in the study of dialogue is to what extent interlocutors converge on shared Linguistic representations, and how they manage to do so (Brennan & Clark, 1996; Pickering & Garrod, 2004). Building on work by Dale and Spivey (2006) and Fernández and Grimm (2014), this study utilizes the relatively novel method of *recurrence quantificational analysis (RQA)* to investigate such linguistic convergence in child-caregiver dialogue. *RQA* (Eckmann, Kamphorst, & Ruelle, 1987) involves the construction of *recurrence plots* – structures which plot two data series against one another, such that every tuple in the coordinate system displays the similarity of two data points. The resultant plots can be visually explored, but they also serve as data structures for the extraction of quantitative measures.

Here, recurrence-plot-derived measures are constructed in order to tease apart the influence of two possible constraints on the extent to which linguistic elements (in our case, words and syntactic structures) are used in both the child's and the caregiver's speech: (1) the prevalence of an element in the other interlocutor's speech and (2) the degree of an element's involvement in temporally close child-adult turns.

In doing so, the study connects to a debate about the nature of child-directed speech (CDS), where a long-standing question is whether CDS results from local adaptation mechanisms or from the caretaker adjusting her speech based on her knowledge of the child's competency (Newport, Gleitman, & Gleitman, 1977; Snow, 1995; Kunert, Fernández, & Zuidema, 2011). In establishing how the adult's speech is affected by the involvement of linguistic elements in temporally close turns, we attempt to measure the impact of local adaptation mechanisms on CDS. If an appreciable impact can be detected, we will be justified in concluding that such mechanisms play a role in the formation of CDS.

In addition to independently measuring the impact of (1a) frequency in the child's speech and (2a) local adaptation mechanisms on the frequency of linguistic elements in the *adult's* speech, we also quantify how the frequency of elements in the *child's* speech is impacted by (1b) their frequency in the adult's speech and (2b) local adaptation. Dialogue is, after all, a coordinative process carried out by two (or more) interlocutors in which all parties are likely to match and influence one another's language use to some extent.

This work thus contributes to ongoing research on adaptation in child-caregiver dialogue by:

- Utilizing and refining RQA for the investigation of adaptation in dialogue.
- Teasing apart and independently assessing the impact of two different factors on the frequency with which words and syntactic structures are used by child and caregiver: (1) their frequency in the other interlocutor's speech; and (2) local adaptation mechanisms.
- Investigating specifically not just how the caregiver adapts to the child's language use, but also how the child adapts to the adult's language use.

Section 2 more carefully defines the notion of *linguistic convergence* as a general kind of local adaptation in dialogue (section 2.1) and reviews the relevant literature on adaptation in child-caregiver dialogue (section 2.2); it furthermore introduces the general methodology (section 2.3) and reviews two imminently relevant studies in some detail (section 2.4). The specific method used in this thesis is presented in section 3. Following that are two analyses – one investigating the impact of local adaptation and frequency in the other interlocutor's speech on child language (section 4), the other investigating how adult language is affected by these two factors (section 5). Results are further discussed in section 6, and section 7 concludes.

2 Background

2.1 Linguistic Convergence

Interlocutors in dialogue match each other's language use through a process of mutual adaptation – that is, they *converge* in their usage of language. Such *linguistic convergence* can be observed at different levels. For example, there exist both Psycholinguistic experiments (Levelt & Kelter, 1982; Weiner & Labov, 1983) and corpus studies (Gries, 2005; Reitter, Moore, & Keller, 2010) which show that people re-use the same syntactic structures shortly after their interlocutors have used them – a phenomenon attributed to *structural priming*. Interlocutors also tend to adopt the same accent, speech rate, and pitch (Giles, Coupland, & Coupland, 1991); they converge on a shared vocabulary – e.g. they tend to use the same or similar spatial descriptions (Garrod & Anderson, 1987) and referential noun phrases (Brennan & Clark, 1996) when carrying out a joint communication-dependent task; and they even match their gaze and postural sway (Shockley, Richardson, & Dale, 2009).

Thus, we use *linguistic convergence* as a cover term for a set of different mechanisms, not all of which may be known, and without committing to the primacy of any one mechanism. By *convergence* we mean simply the process, implemented by some set of mechanisms, which leads to a matching of language use among interlocutors. Possible candidate mechanisms may include priming, repetition, or simply making a dialogue contribution which is relevant to the preceding turn(s). All of these are mechanisms which are unlikely to act on utterances that are far apart in time. That is, a given utterance is not likely to be primed by utterances that are temporally far removed from it; nor, for that matter, is it likely to be a repetition or a causal continuation of such non-local utterances. Hence, when speaking of *convergence*, we always assume a locality-dependence of the underlying mechanisms (evidence for this is discussed in section 2.4).

Processes similar to *linguistic convergence* have been argued to aid in the formation of cognitive representations which facilitate a successful dialogue outcome (Pickering & Garrod, 2004). Zeroing in on a manageable set of lexical items may similarly facilitate successful communication by reducing the number of potentially available words (Brennan & Clark, 1996). In this work, however, the focus is not on why we find *convergence* at so many levels. Rather, we are interested in the effect of *convergence* on the language used by both interlocutors. We refrain, therefore, from appropriating any of the existing nomenclature – such as *alignment* (Pickering & Garrod, 2004) or *entrainment* (Brennan & Clark, 1996) –, in order to highlight the intended theory-independence and generality of the notion of *linguistic convergence*.

2.2 Convergence in Child-Caregiver Dialogue

With respect to *convergence* in child-caregiver dialogue, a large part of the literature focuses on child-directed speech (CDS), a unique kind of speech that adults often use when speaking with children whose linguistic competence is not yet fully developed. Among other things, CDS consists of shorter phrases, contains more pauses, shows a wider range of pitches, and is composed of a limited vocabulary. Since CDS differs so markedly from the kind of speech used in adult-adult dialogue, we may wonder what role it plays in language acquisition (Pinker, 1994; Saxton, 2009). A second question has to do with the mechanisms via which CDS is formed. CDS, to all appearances, results from the caretaker tailoring his or her speech to the child. But is it brought about by local adaptation mechanisms – in other words: *convergence* –, or by the caretaker adjusting her speech based on her knowledge of the child's competency (Newport et al., 1977; Snow, 1995)?

Among other things, we focus on the second question. Put more precisely, the question concerns the manner in which caretakers adjust the level of complexity in their speech as the child matures – a process to which Snow (1989) refers as *finetuning*. Kunert, Fernández, & Zuidema (2011) distinguish between two interpretations of the finetuning hypothesis. On a *weak* interpretation, caretakers rely on knowledge of the child's linguistic proficiency in order to regulate the complexity of their speech. Thus, as the child becomes a more proficient language user, caretakers adjust their speech based on their assessment of the child's proficiency. On a *strong* interpretation of the finetuning hypothesis, *convergence* is responsible for the changing complexity of CDS. That is, the complexity of CDS evolves through largely automatic local mechanisms which operate independently of – but possibly in tandem with – the caretaker's knowledge of the child's linguistic capacities.

One of the earliest studies with a focus on local adaptation in child-caregiver dialogue was conducted by Sokolov (1993), who found that the frequency with which the caretaker applies operations such as *substitution* or *addition* to material from preceding child utterances changes as a function of child age. And since usage of such operations is presumably contingent on the preceding child utterances, and not on the caretaker's knowledge of the child's linguistic competence, Sokolov (1993) interpreted his results to constitute support for local adaptation being involved in the formation of CDS.

This interpretation receives further support through work by Kunert et al. (2011), whose results are strongly suggestive that *convergence* has a part to play in the shaping of CDS. Working with data from the CHILDES corpus (MacWhinney, 2000), they constructed measures representing syntactic, morphological, lexical, and phonological complexity in order to investigate the relationship between child and adult speech. After removing repetitions from the dialogue transcripts, they calculated partial correlation coefficients based on the adult and child values of these measures, controlling for child age. By controlling for age, the caretaker's knowledge of the child's competency – here equated with knowledge of the child's age – is taken out of the picture, making the partial correlations likely to reflect patterns which are caused by *convergence*. And indeed, even after controlling for repetition and child age, Kunert et al. (2011) found a weak correlation for an aggregate measure of linguistic complexity (none of the individual measures led to significant results). This finding suggests that there must be something other than the caretaker's knowledge of the child's competence to explain the correlation between the linguistic complexity of the child's and the adult's speech, with *convergence* being a prime explanatory candidate. Their methodology, however, did not allow for further exploration of this possibility.

In this study, we aim to develop and apply a method which can more directly measure the effect of *convergence* on the adult's language use. We achieve this by deriving measures which track (1) the frequency of words and syntactic structures in the child's speech and, separately from that, (2) the extent to which words and syntactic structures are involved in local adaptation mechanisms. This allows us to measure the impact of (1) and (2) on the frequency of words and syntactic structures in the adult's speech. And if we can detect an appreciable impact of (2), we will be justified in concluding that CDS is – at least in part – the result of local adaptation mechanisms. Beyond that, by comparing the impact of (1) to that of (2), we may also gain insight into the relative importance of frequency and *convergence* for the formation of CDS.

But moving beyond CDS, we also consider the role of *convergence* and frequency in the shaping of the *child's* speech. It is well-established that caretaker and child speech are closely related in terms of their global properties. For example, CDS and child speech have been found to show similarities with respect to verb and noun proportions (Choi & Gopnik, 1995; Tardif, Shatz, & Naigles, 1997) as well as with respect to inflectional patterns (Aksu-Koç, 2006) and specific constructions, such as *determiner-noun* or *numeral-noun* (Cameron-

Faulkner, Lieven, & Tomasello, 2003). We should, therefore, certainly expect the frequency of elements in CDS and their frequency in the child's speech to affect one another.

What is less clear is whether the involvement of linguistic elements in *convergence* has any effect on their usage in the child's speech. The work discussed above by Sokolov (1993) and Kunert et al. (2011) suggests that *convergence* is a measurable property of child-caregiver dialogue, which is likely involved in the shaping of CDS. But of course we should not expect *convergence* to only affect CDS; it probably exerts an influence on the child's speech as well. And to date, we do not find much work in this direction. A notable exception are Veneziano & Parisse (2010), who examined the effect of local adaptation on the frequency with which French-speaking children use the phonological forms of inflected verbs.

In French, several grammatical distinctions marked in writing are not reflected in the pronunciation of verbs – although compared to English, spoken French still contains a large number of phonologically distinct morphological verb forms (phonological verb forms). This has consequences for the acquisition of verbal morphology in that children will initially use only one phonological verb form, in effect ignoring some of the grammatical distinctions reflected in French phonological verb morphology. Eventually, children begin to use the full array of phonological verb forms, but they do so only after a period of using just a single form. Crucially, which phonological verb form is the one first acquired may very well depend on its frequency in CDS *and* on the extent to which it is involved in *convergence*.

For each of the verbs associated with several phonologically distinct forms, Veneziano & Parisse (2010) focused on the timespan within which the two children observed by them (ages 1;31 – 2;26 and 1;7.2 – 2;3.4) used the verb in just a single phonological form. They then classified this phonological form as being (1) predominantly influenced by CDS, (2) predominantly influenced by *convergence* (*conversational contingencies*, in their terminology), (3) equally influenced by both, or (4) influenced by neither. Each phonological verb form was assigned to one of the four classes based on how often it occurred in CDS (measuring the influence of CDS on usage) and how often the child used it immediately after or immediately before the adult's usage of the same form (measuring the influence of *convergence* on usage).

For example, if at least 75 % of the occurrences of a given verb in CDS corresponded to the phonological form used by the child, that form was taken to be very strongly influenced by CDS.[1] And if the child's usage of the form was immediately followed or preceded by the adult's usage of the same form for at least 75 % of its occurrences in the child's speech, it was additionally considered to be very strongly influenced by *convergence*. In this example, since the influence of both CDS and *convergence* is considered to be very strong, the phonological verb form used by the child would be classified as *equally influenced by both*. The situation would be different if the child's usage of the phonological form was followed or preceded by the same form, produced by the adult, for less than 75 % of its occurrences in the child's speech. Assuming that the form's frequency in CDS remained unchanged, Veneziano & Parisse (2010) would no longer attest a very strong influence of *convergence*, and the form would be classified as *predominantly influenced by CDS*.

By defining a number of different frequency thresholds for the strength of the influence of CDS and *convergence* on a given phonological verb form, Veneziano & Parisse (2010) determined which of the two, if any, is the more dominant influence. Using this procedure, each phonological verb form used by the child was assigned to one of the four classes from above. It emerged that about half the phonological verb forms fell into category (3) – i.e., they were equally strongly influenced by CDS and *convergence*. Importantly, about three quarters of the other half fell into category (2) – meaning *convergence* was the more important factor in making them the dominant forms in the child's speech.

[1] But note that Veneziano & Parisse (2010) do not use the term *very strongly influenced* – the vocabulary used to convey the general ideas behind their design has been adapted for this short exposition.

Veneziano & Parisse's (2010) results show that *convergence* probably has a part to play in shaping the child's usage of language. But in distinguishing general frequency in CDS from usage in *convergence*, their study also introduced an important methodological innovation, which we attempt to incorporate in the design of the present study. There, we combine this general idea – to distinguish between *convergence* and frequency in the other interlocutor's speech – with a more sophisticated method, one which is uniquely suited for measuring local adaptation processes in dialogue. The method in question is called *Recurrence Quantificational Analysis (RQA)* and has been used to show that conversational units in close temporal proximity are subject to lexical, syntactic, and semantic adaptation (Dale & Spivey, 2006; Fernández & Grimm, 2014). The following section 2.3 introduces the basic method, and section 2.4 discusses two studies which have utilized *RQA* to measure *convergence* in child-caregiver dialogue.

2.3 Recurrence Plots for Dialogue

Introduced by Eckmann et al. (1987) for the graphical and quantitative exploration of the behavior of evolving systems, *RQA* is typically used in data-heavy fields such Physics and the Earth Sciences (Marwan & Kurths, 2002; Gao & Cai, 2000), while relatively few investigators have used the technique for language research (Dale & Spivey, 2005, 2006; Angus, Smith, & Wiles, 2012; Fernández & Grimm, 2014).[2] The technique relies on the construction of so-called *recurrence plots*, which serve both as visualization tools for complex data series and as data structures for the extraction of quantitative measures. To construct a recurrence plot, one minimally requires two time series of data, one plotted on the x-axis and the other plotted on the y-axis.

In the case of dialogue, a series might consist of all conversational turns by a given interlocutor, indexed by time so that turn 1 temporally precedes turn $1 + n$.[3] Assuming that we are dealing with two interlocutors, A and B, we might plot A's turns on the x-axis and B's turns on the y-axis.[4] Every point in the coordinate system of the recurrence plot then corresponds to two turns – one by A and one by B. We now wish for coordinates in the plot to display information about the similarity of their data points; continuing with the example of turns, one possible way to achieve this would be to assign a similarity score of 1 to the coordinate if one of the two turns being compared is an exact lexical repetition of the other; else, we assign a similarity score of 0. Such a Boolean plot – easily visualized by coloring coordinates black if their value is 1 and white otherwise – would then convey information about the overlap in the turns produced by the two speakers.

For example, suppose we wish to apply this method to the following eight-turn dialogue:

A_0: Aren't recurrence plots fascinating?

B_0: I don't find them very fascinating, to be honest.

A_1: Really?

B_1: There are more interesting topics...

A_2: There are more interesting topics!

[2]For an introduction to recurrence plots, see DiDonato et al. (2013); a review of recurrence plots as used in the analysis of conversation is given by Leonardi (2012).

[3]Here and elsewhere in the thesis, we consider any uninterrupted stretch of speech produced by one speaker to be a *turn*

[4]Technically, what we are describing here is a *cross-recurrence plot*, which is a type of recurrence plot where different data series are plotted against each other, rather than plotting the same series on each axis. In this thesis, the term *recurrence plot* is used interchangeably with the term *cross-recurrence plot*.

B_2: Don't act so outraged now. It's just that I prefer to do theoretical linguistics.

A_3: Okay, that's fair enough. You're a theoretically-minded person.

B_3: Exactly!

In this toy dialogue, speaker A and speaker B contribute a series of four turns each – $< A_0, A_1, A_2, A_3 >$ for A and $< B_0, B_1, B_2, B_3 >$ for B. If we place A's turns on the x-axis and B's turns on the y-axis, we end up with a recurrence plot of size 4×4. Since turn A_2 is an exact lexical repetition of turn B_1, the coordinate $(2,1)$ is assigned a similarity score of 1.[5] And since there are no other exact repetitions within the dialogue, all other coordinates are assigned a score of 0. We can visualize the plot as a grid of tiles, where every tile is associated with a coordinate. We then color the tile corresponding to coordinate $(2,1)$ black and all other tiles white. The result is shown in Figure 1.

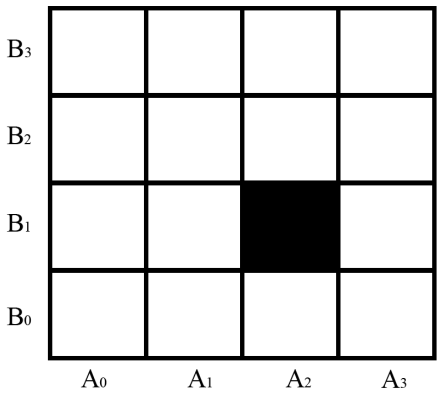

Figure 1: Turn-based recurrence plot for a toy dialogue consisting of eight turns by two speakers. Turns $A2$ and $B1$ are assigned a similarity score of 1, which is visualized by coloring the corresponding tile at coordinate $(2,1)$ black. The remaining coordinates correspond to pairs of turns whose similarity score is 0, and so their tiles are colored white.

Observe that turns in close proximity are located on a *diagonal line of incidence*. In the toy dialogue, we find seven pairs of turns which are immediately adjacent – (A_0, B_0), (A_1, B_0), (A_1, B_1), (A_2, B_1), (A_2, B_2), (A_3, B_2) and (A_3, B_3). The coordinates corresponding to these pairs are located on a diagonal line, going from the lower left to the upper right corner of the recurrence plot. This is illustrated in Figure 2, where tiles corresponding to immediately adjacent turns are colored in gray.

In more precise terms, the *diagonal line of incidence* will always include coordinates where the time index i of speaker A is equal to the time index j of speaker B. In addition, it will also include either points $(i+1, i)$ or $(i-1, i)$, depending on whether A or B, respectively, has uttered the first turn in the dialogue. In other words, the *diagonal line of incidence* always coincides with the true diagonal, going from the lower left to the upper right corner of the recurrence plot, plus the coordinates either immediately above or below the true diagonal – depending on which of the two speakers initiated the dialogue. In the plot based on the example dialogue between A and B, the *diagonal line of incidence* contains points on the true diagonal – the coordinates $(0,0)$, $(1,1)$,$(2,2)$, and $(3,3)$ –, plus a series of points immediately below the true diagonal – i.e., the coordinates $(1,0)$, $(2,1)$, and $(3,2)$. If B and not A had begun the dialogue, this last series would consist of points immediately above the true diagonal – i.e., the coordinates $(0,1)$, $(1,2)$, and $(2,3)$.

[5] Assuming, that is, that the first turns in each series are indexed with 0 – a convention followed throughout.

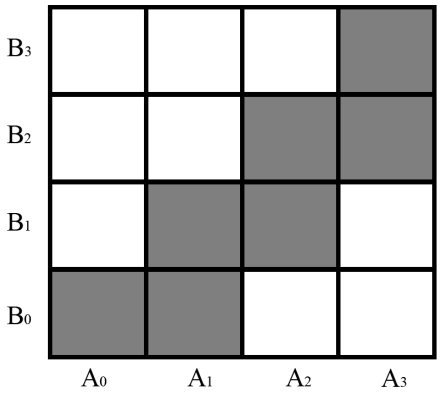

Figure 2: Recurrence plots with tiles falling on the *diagonal line of incidence* colored in gray. Turns compared at coordinates on the *diagonal line of incidence* are immediately adjacent.

The situation becomes more complicated if one interlocutor contributes more data points than the other, in which case the *diagonal line of incidence* may be quite far removed from the true diagonal. But in dialogues with two participants, the difference between the numbers of turns per interlocutor is either zero or one, and so turn-based recurrence plots are easily made quadratic by removing the last turn of the interlocutor with the greater number of turns. The advantage of such an approach is that the diagonal line of incidence can easily be identified via a straightforward algorithmic procedure, whose only parameter is who of the two speakers contributed the first turn. Because of this, we will always ignore the last turn of the speaker with the greater number of turns. For our purposes, that is, both interlocutors always contribute an equal number of data points.[6]

Given a (quadratic) recurrence plot, perhaps the most coarse-grained measure we can extract from it is the global recurrence rate *RR*, which is the sum of all similarity values in the plot divided by the total number of coordinates. Formally:

$$\text{RR} \quad = \quad \frac{\sum_{i,j \leq n} m(i,j)}{n \times n}$$

where m is some similarity function, n is the length of a single speaker's data series (equal to the length of the other speaker's data series), i is the time index of speaker A, and j is the time index of speaker B. This measure allows us to quantify the average recurrence within the entire dialogue, but it ignores temporal proximity.

Luckily, temporal information is implicitly encoded within a recurrence plot, in that data points compared at coordinates on the *diagonal line of incidence* are immediately adjacent; and the farther away a given coordinate is from the diagonal, the larger the temporal distance between the two data points compared at that coordinate. Equipped with this insight, we define a distance d of data points above and below the diagonal which we will include in the calculation of the local recurrence rate RR_d. For example, if $d = 3$, we will include points on the diagonal as well as points up to three coordinates removed from it. To obtain RR_d, we take the sum of all similarity scores no more than d points removed from the diagonal and divide the result by the total number of coordinates involved in the summation. Formally (cf. Fernández & Grimm, 2014, p. 2):

$$\text{RR}_d \quad = \quad \frac{\sum_{i \leq n} \sum_{j \in [i-d, i+d]} m(i,j)}{|D|}$$

[6]But cf. Dale & Spivey (2006, p. 405) for a way to estimate a hypothetical *diagonal line of incidence* in cases where the two data series are not of equal length.

where D is the set of all points on the *diagonal line of incidence*, plus points no more than d coordinates removed from it.[7] Because coordinates that are farther away from the *diagonal line of incidence* will compare data points with a larger temporal distance between them, a large d will define an area around the diagonal that includes more similarity scores which are based on pairs of distant data points; accordingly, the calculation of RR_d will involve a higher number of such scores. For recurrence plots based on linguistic similarity measures, we typically find that as we increase d, the average recurrence rate decreases, meaning that e.g. turns in close temporal proximity tend to be more similar than turns which are farther apart.

It is furthermore possible to break down RR_d into the *A-initiated recurrence rate* and the *B-initiated recurrence rate* – or in the case of child-caregiver dialogue, *adult-initiated recurrence rate* and *child-initiated recurrence rate*. We can calculate the former by adding up all the coordinates in a recurrence plot where the child speaks after the adult, then dividing the result by the number of data points involved in the addition; and we can obtain the latter by doing the same for all coordinates where the adult's dialogue contribution follows the child's contribution. The *adult-initiated recurrence rate* then corresponds to that part of the overall recurrence rate which is due to the child adapting his or her language use to that of the adult, while the *child-initiated recurrence rate* represents the part which is due to the adult adapting to the child's usage.

2.4 Dale & Spivey (2006) and Fernández & Grimm (2014)

Dale & Spivey (2006), who analyzed child-caregiver dialogue from the CHILDES database, were the first to apply *RQA* to child-caregiver dialogue. Instead of using turns, as in the example discussed above, their recurrence plots are based on part-of-speech-tag bigrams (POS bigrams). That is, instead of sequences of turns, their data series consisted of pairs of POS tags. Consider a simple example dialogue, where A initiates the exchange by saying, "Recurrence is interesting," to which B responds, "It *is* interesting!" We might then have the following POS tag sequences: *noun, auxiliary, adjective* for A and *pronoun, auxiliary, adjective* for B. Taking a POS bigram to be a pair of POS tags, we would obtain the following two data series: *(noun, auxiliary), (auxiliary, adjective)* for A's turn and *(pronoun, auxiliary), (auxiliary, adjective)* for B's turn. If we plot the two bigram series against one other, we obtain a recurrence plot of size 2×2, where the only equivalent POS bigrams are the second in A's series and the second in B's series. Hence, we assign a similarity score of 1 to coordinate $(1,1)$, while the remaining three coordinates are assigned a score of 0.

Dale & Spivey (2006) applied this procedure to all dialogues in three corpora from the CHILDES database (Sarah from the Brown corpus, Abe from the Kuczaj corpus, and Naomi from the Sachs corpus). For each of the three corpora, they calculated as many recurrence rates as there are dialogue transcripts in the corpus – one for each dialogue. They then took the mean of the recurrence rates for each corpus, ending up with a single average recurrence rate per corpus. By repeating this for different values of d, they calculated a set of average recurrence rates – each for a different level of locality d. And for this set of average local recurrence rates, we can see clearly that as we increase d, the recurrence rate decreases.

Importantly, Dale & Spivey (2006) also found that for small values of d, the average recurrence rate differs significantly from what would be expected by chance. In one of their two baseline conditions, they constructed recurrence plots based on randomized transcripts, where the utterances of the child where randomly shuffled, while the order of adult utterances was kept intact;[8] and given a relatively small d, the average recurrence rate is signifi-

[7] Note that RR is a special case of RR_d, namely $RR = RR_d$ for $d = n$.

[8] In their second baseline condition, utterances by the caregiver were replaced with his or her utterances

cantly different from this shuffled baseline recurrence rate. It is also this pattern – a higher recurrence rate for temporally close turns, statistically different from a randomized baseline – which adds support to the assumption that *linguistic convergence* is a local phenomenon. In other words, we assume that we find an increased similarity of temporally close conversational units *because* of the convergence process discussed in sections 2.1 and 2.2.

As far as the contrast between *adult-initiated recurrence rate* and *child-initiated recurrence rate* is concerned,[9] Dale & Spivey (2006) distinguished only between global *adult-initiated recurrence rate* and *child-initiated recurrence rate* – i.e. the level of locality d was as large as possible. By comparing the two rates, they were able to infer who of the two interlocutors adapts more to the other. If the *child-initiated recurrence rate* is larger, the child can be seen to lead the adaptation process since the adult is adapting more to the child than the child to the adult. Otherwise, if the *adult-initiated recurrence rate* is larger, the child adapts more to the adult. Here, Dale & Spivey's (2006) finding was that although the overall recurrence rate is rather evenly split among both interlocutors, in some dyads the child adapts more to the adult, while in other dyads the adult adapts more to the child.

Introducing some modifications and complications into their method, Fernández & Grimm (2014) measured the similarity of conversational turns in the same corpora of child-adult dialogue used by Dale & Spivey (2006) and additionally considered adult-adult dialogue. In their turn-based plots, all turns by speaker A are placed on the x-axis, and all turns by speaker B are placed on the y-axis. Every coordinate then corresponds to a similarity score for a pair of turns, where scores can take any value on a continuum between 0.0 and 1.0, meaning that the plots contain graded rather than Boolean 1 or 0 values. Following work by Angus et al. (2012), coordinates are colored black for maximal similarity, white for minimal similarity, and in shades of grey for intermediate values. Within the plots, immediately adjacent turns are compared along the diagonal; and because such turns are usually more similar than turns with a greater temporal distance, plots often have a visibly darker *diagonal line of incidence.* Figure 3 shows two examples.

From a dialogue perspective, the approach taken by Fernández & Grimm (2014) makes the recurrence plots more intuitively appealing and avoids difficulties in estimating the *diagonal line of recurrence* by making turns the units of comparison. As discussed in the previous section, in dialogues with two participants, the difference between the number of turns per speaker is always either zero or one; and by removing the last turn of the interlocutor with the greater number of turns, recurrence plots are kept quadratic. This simplifies things in that it is much more straightforward to identify the *diagonal line of incidence*, which will always consist of coordinates on the true diagonal – i.e., the line from the lower left corner going to the upper right corner of the plot –, plus the points immediately above or below the true diagonal, depending on whether the child or the adult contributes the first turn.

In addition, Fernández & Grimm (2014) determined similarity using more diverse measures. The first of their three similarity measures calculates the similarity of any two turns in terms of their shared lexemes, where lexemes are pairs of stems and POS tags (e.g. *(cat, noun)*). This measure thus tracks lexical similarity. The second measure – closest to the one

from one transcript ahead. Since results for the two baselines often are not very different, we limit discussion to the shuffled baseline.

[9] Dale & Spivey (2006) did not use the terms *adult-initiated recurrence rate* and *child-initiated recurrence rate*. In their work, the former is denoted by $RR+$, and the latter is denoted by $RR-$. The reason is that if the child's coordinates are plotted on the y-axis, and the adult's points are plotted in the x-axis, all coordinates above the diagonal line of incidence are such that the child speaks after the adult, and all coordinates below the diagonal are such that the adult speaks after the child. The plus sign thus signals that points above the *diagonal line of incidence* count towards $RR+$, and the minus sign signals that points below the diagonal count towards $RR-$. Perhaps at the cost of some geometrical clarity, we use the more intuitive terms *adult-initiated recurrence rate* and *child-initiated recurrence rate* instead.

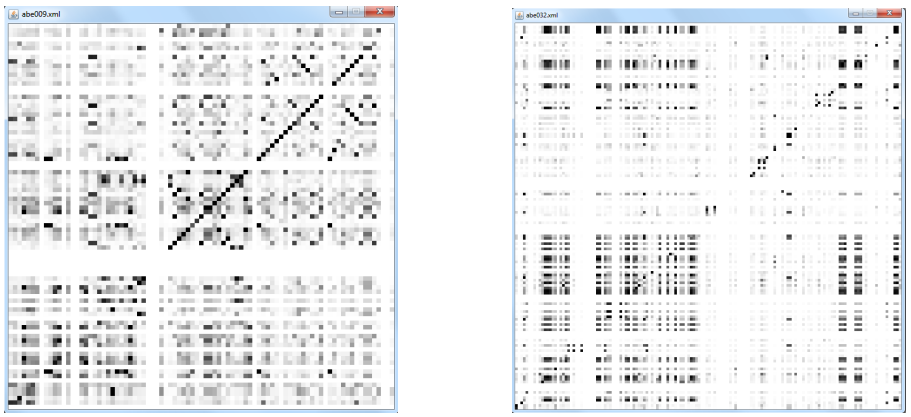

Figure 3: Recurrence plots from dialogues in the Kuczaj folder (CHILDES database). Plots are based on a distributional semantic similarity measure. Left: Kuczaj folder, child age 2;5.26. Right: Kuczaj folder, child age 2;8.22.

used by Dale & Spivey (2006) – determines similarity based on shared POS bigrams, where a POS bigram is considered to be shared by two turns just in case both turns include it *and* at least one of the two terminal nodes associated with the POS tags is not the same in both turns. Since it is based on POS bigrams, this second measure takes the ordering of words into consideration, which makes it more syntactic in nature. And by requiring shared POS bigrams to differ in terms of at least one terminal node, it further emphasizes abstract syntactic over lexical similarity. Finally, the third and last of Fernández & Grimm's (2014) measures is based on a distributional semantic model. Such models derive vector-representations of words that capture the meaning of a single word in terms of the contexts within which it typically occurs. This measure constitutes an attempt to track not just the similarity in meaning of equivalent words – any word is obviously semantically similar to itself – but also the meaning-related similarity of words such as *water* and *river*, or *fork* and *dinner*. These words are dissimilar on a string level, even though they are more or less closely related in terms of their meaning.

The upshot of Fernández & Grimm's (2014) work is that also for their three similarity measures, we observe the effect that smaller values of d lead to a larger recurrence rate within child-caregiver dialogue. In other words, turns in close temporal proximity are more lexically, syntactically, and semantically similar than turns which are farther apart.[10] And just as reported by Dale & Spivey (2006), as we increase d, the recurrence rate approaches the randomized baseline level. Figure 4 below plots selected results.

Fernández & Grimm's (2014) improved method thus appears to produce reasonable results in that it was successfully used to identify the same general pattern reported by Dale & Spivey (2006). It is noteworthy that this pattern emerged not just for the syntactic similarity measure, but also for the lexical and semantic measures. We may conclude, then, that Fernández & Grimm's (2014) method is suitable for tracking different types of *linguistic convergence*. It is, therefore, reasonable to adapt their refined method to measure the extent to which *linguistic convergence* affects the frequency with which child and adult use specific linguistic elements. As we aim to measure the involvement of *single* linguistic elements in convergence, we will determine the similarity of pairs of turns in terms of the presence or ab-

[10]Surprisingly, while Fernández & Grimm (2014) found this to be the case also for the lexical and semantic similarity of turns in adult-adult dialogue from the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992), they discovered the opposite pattern for syntactic similarity. For their *POS bigrams* measure, the recurrence rate in adult-adult dialogues from the Switchboard corpus is actually *smaller* for small values of d, and *less* than what would be expected by chance. But however surprising, the present study does not deal with adult-adult dialogue, and so we will not further discuss this finding here.

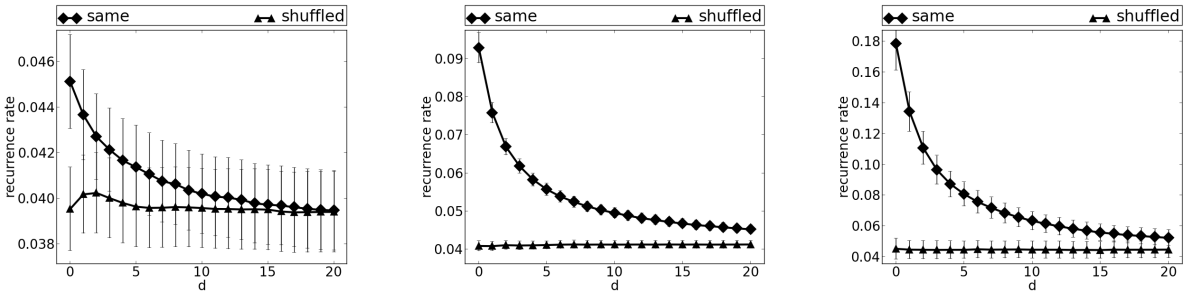

Figure 4: Average recurrence rate as a function of the level of locality d, plotted based on results obtained by Fernández & Grimm (2014). For small values of d, the average recurrence rate is significantly larger than its counterpart in the baseline condition, where the child's turns are randomly shuffled. As we increase d, result for the experimental condition approach results for the shuffled baseline condition. Left: *POS Bigrams*; Sarah corpus, Brown folder. Middle: *Lexemes*; Abe corpus, Kuczaj folder. Right: *Distributional Semantic Model*; Naomi corpus, Sachs folder.

sence of just a *single* linguistic element; and to simplify the procedure, we will use Boolean rather than graded recurrence plots. But even with these changes, Fernández & Grimm's (2014) recurrence-quantificational tools still serve as the basis for the methodology that is expounded in the following section.

3 Method

3.1 Outline

The goal of this study is to measure the extent to which adult and child usage of linguistic elements is affected by two separate influences: (1) the extent of an element's involvement in *convergence*, and (2) the frequency with which an element occurs in the other interlocutor's speech. We consider a *linguistic element* to be a component unit of a turn, at any level of linguistic processing (e.g. a lexeme, morpheme, or phoneme). The notion of *linguistic convergence* is then defined as follows:

> **Linguistic convergence:** A process whereby interlocutors in dialogue converge on a shared set of linguistic elements which comprise their turns. *Convergence* here subsumes different mechanisms such as (structural) priming, imitation, or other forms of repetition, but also simply making dialogue contributions which are relevant to the preceding turn(s). These mechanisms all act in unison to produce an overall convergence of linguistic elements, at different linguistic levels. And while it may be plausible that one or more of the underlying mechanisms act as the primary cause of convergence, we remain agnostic with respect to which individual process, if any, is the dominant force. We do, however, assume a locality-dependence of the underlying mechanisms, in that a given turn will rarely be primed by, be a repetition of, or causally connect to a turn which is temporally far removed from it. That is, only turns in relatively close temporal proximity are subject to *linguistic convergence*.

A crucial assumption is that *convergence* can be measured using *RQA*. In particular, we assume it can be measured by the recurrence rate for small values of d – which essentially measures the overlap in linguistic elements of conversational turns within close temporal

proximity.[11] As illustrated in the previous section, such turns show a higher overlap than what would be expected by chance (cf. Figure 4), and it is precisely this non-random overlap which is likely caused by a process of *linguistic convergence*.

In the present study, we consider three different types of linguistic elements: content words, function words, and part-of-speech tag bigrams (POS bigrams). We split the general category *words* into the two sub-categories *content words* and *function words* to measure (1) a more meaning-driven usage (content words) and (2) a more syntactically oriented or stylistic usage of lexical items (function words). We also find some precedence in the literature for considering the role of function words separately from the role of content words. For example, the coordination of function words has been used to measure *language style matching*, which has been found to be predictive of success in communication-dependent shared tasks (Gonzales, Hancock, & Pennebaker, 2010) and of future romantic entanglement between language-style-matching interlocutors (Ireland et al., 2011). Danescu-Niculescu-Mizil & Lee (2011) even found such matching in the fictional dialogue of movie scripts, which they take to indicate that the matching process is mostly subconscious. Adaptation involving function words may thus reflect a syntactically oriented type of style-convergence, whereas adaptation involving content words may reflect a more semantically oriented kind of convergence.

Lastly, the third category, *POS bigrams*, has been designed to capture more purely syntactic *convergence*. In fact, we assume that POS bigrams stand in rough correspondence to syntactic trees. By looking not at single words but at sequences of two POS tags, we take into account the ordering of words, where the words themselves may or may not be equivalent across any two converging turns, so long as the POS tags are identical. Hence, by emphasizing the order of words and by de-emphasizing lexical repetition, this category is surely more syntactic in nature than the other two measures. And being the most syntactic out of the three measures, it marks the end-point in a continuum from semantic (content words), to more syntactic (function words), to yet more syntactic in nature (POS bigrams).

Using corpus data from the CHILDES database, we perform two different analyses for each of the three element types. The first analysis teases apart and independently measures the impact of (1) frequency of linguistic elements in the adult's speech and (2) *convergence* on the frequency of linguistic elements in the child's speech. It is only in the second analysis that we connect to the debate about CDS. There, we focus on the adult's speech, i.e. we measure the impact of (1) frequency in the child's speech and (2) *convergence* on frequency in the adult's speech.

For each of the two analyses, we additionally break down *convergence* into two separate measures: one which tracks occurrences of linguistic elements used in *convergence* where the adult uses an element after the child has used it; and one which tracks occurrences in *convergence* where the child's usage of an element follows the adult's usage of the same element. Using these measure, we are able to determine if *convergence* that is due to the adult adapting to the child has a stronger impact on child / adult frequency than *convergence* that is due to the child adapting to the adult.

3.2 Corpus Data

The data for this study come from three English corpora in the CHILDES database (MacWhinney, 2000): Abe from the Kuczaj folder (Kuczaj & Stan, 1977); and Adam and Sarah from the Brown folder (Brown, 1973). Each corpus consists of transcripts of verbal in-

[11] Although, to be more precise, this depends on the similarity measure used. But since most of the similarity measures discussed, and all of the measures used in this study, determine the similarity of pairs of turns in terms of the overlap with respect to particular linguistic elements, we will not further elaborate on this point.

teraction between a US-American child – Abe, Sarah, and Adam, respectively – and at least one adult, who typically is one of the child's parents, although the number of adults present and their relation to the child differs from transcript to transcript. The transcripts are based on recordings which are more or less evenly spaced over a period of 20 – 32 months. For the most part, the recording was carried out in the homes of the children, at different times of day.

Since the method used in this study relies on plentiful data to produce statistically significant results, the corpora were chosen so that they would yield both a relatively large number of dialogues and a relatively large number of word tokens. Because of this, the data also cover a relatively wide age range. Table 1 displays for each corpus the age ranges covered, the total number of word types and tokens, the total number of dialogues, and the average number of turns per dialogue. Note that transcripts where the child's utterances have a mean length of less than two words were not considered for analysis.

Corpus	Age Range	Nr. Word Tokens	Nr. Word Types	Nr. Dialogues	Avg. Nr. Turns / Dialogue
Abe	2;4 – 5;0	282.964	4.630	210	191 (sd: 74)
Sarah	2;3 – 5;1	231.940	4.834	107	340 (sd: 84)
Adam	2;3 – 4;1	273.994	4.041	54	759 (sd: 148)

Table 1: Details of corpora used in analysis, minus transcripts where the child's utterances have a mean length of less than two words. The numbers for *Word Tokens*, *Word Types*, and *Dialogues* are totals for the whole corpus.

3.3 Recurrence Quantificational Measures

For the analyses, we compute various recurrence-quantificational measures, based on the following three types of linguistic elements:

- **Content Words:** pairs of stems and part-of-speech tags (POS tags), where the tag is either a noun, adjective, or verb (e.g. *(cat, noun)*, or *(move, verb)*).
- **Function Words:** pairs of stems and POS tags, where the tag is either the infinitive marker *to*, an adverb, pronoun, conjunction, negation particle, auxiliary verb, modal verb, preposition, quantifier, modifier, or determiner (e.g. *(they, pronoun)*, or *(the, determiner)*).
- **POS Bigrams:** pairs of POS tags, where the first tag belongs to a terminal node which immediately precedes the terminal node associated with the second tag – e.g. *(determiner, noun)*, or *(verb, noun)*. POS bigrams are assumed to correspond roughly to syntactic structures.

The next step is the calculation of measures which tease apart and independently assess the impact of general frequency and *convergence* on the usage of the above kinds of elements. To do this, we pick a suitable linguistic element E and construct a turn-based Boolean recurrence plot where two turns are given a similarity score of 1 just in case both the child and adult turn contain E, and 0 otherwise. This plot can then aid us in calculating the sought-after measures, a list of which is given in Table 2 below. Bold-faced items are measures which are used for analysis, while non-bold-faced items represent intermediate steps in the calculation of the final measures.[12]

[12]It may be of interest that while the measures in Table 2 are all related to the various types of recurrence rates discussed in sections 2.2 and 2.3, none of them actually involve recurrence *rates*. That is – they only ever

Measure	Definition	Symbol / Formula
Recurrence	Given a dialogue and some linguistic element, construct a recurrence plot for the element based on the dialogue. Given an additional distance d from the diagonal line of incidence, the *Recurrence* is the sum of scores within the area defined by d around the diagonal.	R_d
Random Recurrence	The *Recurrence* for a recurrence plot where the child turns have been randomly shuffled. Corresponds to the amount of recurrence that would be expected by chance and is used as a baseline.	R_d^{ran}
Child-Initiated Recurrence	Given a recurrence plot based on some dialogue and a distance d from the diagonal line of incidence, the sum of scores within the area defined by d, for all pairs of turns where the adult turn follows the child turn.	R_d^c
Adult-Initiated Recurrence	This measure is the same as *Child-Initiated Recurrence*, except that we consider only pairs of turns where the child turn follows the adult turn.	R_d^a
Adult / Child Frequency	Given a dialogue and some linguistic element, the frequency with which that element is used by the adult or child within the dialogue.	$F_{a/c}$
True Adult / Child Occurrence	Given a dialogue, some linguistic element, and a distance d, construct a recurrence plot for the dialogue. This measure is then the frequency with which the element is used by the child or adult, minus occurrences of the element which fall within the area around the diagonal defined by d on the recurrence plot. It measures the frequency of an element without considering occurrences that count towards *Recurrence*.	$O_d^{a/c}$
True Recurrence	*Recurrence* divided by the sum of adult and child frequency. If the resultant value is high, the linguistic element in question has a high chance of recurring – regardless of its frequency in child and adult speech.	$\frac{R_d}{F_a+F_c}$
True Adult-Init. Recurrence	This measure is the same as *True Recurrence*, except that it is only based on pairs of turns where the adult turn follows the child turn. Thus, the higher its value, the higher the chance that the adult will use the linguistic element in question within d turns after the child has used it.	$\frac{R_d^a}{F_a+F_c}$
True Child-Init. Recurrence	Same as *True Recurrence*, except that we only consider pairs of turns where the child turn follows the adult turn. Therefore, a high value signifies a high chance that the child will use the element in question within d turns after the adult has used it.	$\frac{R_d^c}{F_a+F_c}$

Table 2: Measures used in analysis. Non-bold-faced items represent intermediate steps in the calculation of the final, bold-faced measures.

With this set of recurrence-quantificational measures at our disposal, we proceed as follows. First, we must choose a value for the distance d from the *diagonal line of incidence*, which is required for the calculation of those measures which involve recurrence plots. Since we have established that *convergence* is locality-dependent (cf. the discussion in section 2.4 about smaller values of d giving rise to large recurrence rates), we should pick a relatively small d. We decide to set $d = 2$, while noting at the same time that the decision remains somewhat arbitrary, in that other relatively small values would also be suitable.

Given the linguistic element E, we use $d = 2$ to calculate the various measures. Importantly, we do this for every dialogue in the corpus, meaning that we end up with several measures of the same type. For example, if there are 100 dialogues in the corpus, we will have calculated 100 *Child / Adult Frequencies*, 100 *True Child / Adult Occurrence* scores, 100 *True Recurrence* scores, and so on. At that point, we take the average of every score, collapsing the different dialogues into one value per measure. This move serves two practical purposes: for one, the patterns of convergence become most apparent when generalizing over a relatively large number of dialogues; and for another, by taking this more coarse-grained approach, we abstract away from the details, and we are likely to discover robust, general patterns – an approach befitting the exploratory nature of this work.

As a requirement on the averages, we would like them to be significantly different from what would be expected by chance. To test this, we calculate the average *Random Recurrence*, which is the average *Recurrence* with the child's turns randomly shuffled. As *convergence* is a local phenomenon, we find no significant difference between *Recurrence* and *Random Recurrence* for large values of d, across linguistic elements. That is: the *Recurrence* we get for large values of d is due to chance – it is only the *Recurrence* we get for small values of d that is due to a non-random process of *linguistic convergence*. Thus, we only consider E for analysis if there is a significant difference between the average *Random Recurrence* and the average *Recurrence* at $d = 2$.[13] In this fashion, we calculate average measures for all linguistic elements that show such a significant difference. Accordingly, in the remainder of this thesis, when we speak of e.g. the *True Recurrence* of some E, we mean its average *True Recurrence*.

3.4 Procedure

In order to independently assess and compare the impact of (1) general frequency and (2) *convergence* on the frequency with which linguistic elements are used by the child and adult, we focus predominantly on *True Occurrence* and *True Recurrence*. *True Recurrence* tracks the extent to which an element is used in *convergence*, independently of its general frequency – it is a frequency-independent measure of *convergence*; and *True Occurrence* tracks the frequency of elements but factors out occurrences that are used in *convergence* – i.e., it is a *convergence*-independent measure of frequency. The idea is then to use these two measures to predict both *Child Frequency* and *Adult Frequency*.

To achieve this, we make use of multiple linear regression analysis. A multiple linear regression models the relationship between two or more predictor variables and a response variable; and the slopes or coefficients associated with the predictor variables convey infor-

involve a sum of similarity scores, but these scores are not divided by the total number of coordinates. Division by the total number of coordinates may be useful for showing that quotients for smaller values of d are larger than those for larger values of d. But this fact is not exploited in the design of this study. So for simplicity, we operate with sums of similarity scores, rather than normalizing those sums by the number of similarity scores involved in the summation.

[13] One consequence of this is that in addition to the theoretical reason for choosing a small d (i.e. because *convergence* is a local phenomenon), we now also have a practical reason. Namely, if we picked too large a d, there would not be enough linguistic elements that show a significant difference between the average *Random Recurrence* and the average *Recurrence* to perform a meaningful analysis.

mation about the average increase in the response variable, given an increase in the predictors. Within *multiple* linear regression models, these are *partial* regression coefficients, which factor out the increase in the response variable that is caused by the other predictors. Thus, multiple linear regressions control for possible correlations of the predictors, preventing a scenario where we ascribe the same effect to both variables. For example, say we model the growth of *Child Frequency* using the two predictors *True Adult Occurrence* and *True Recurrence*. Assume, furthermore, a statistically significant value of 0.8 for the *True Adult Occurrence* coefficient. In such a scenario, if *True Adult Occurrence* increases by 1, then the average increase in *Child Frequency* will be 0.8, given that *True Recurrence* is held constant. In other words, no part of this 0.8 increase is also caused by an increase in *True Recurrence*.

In constructing the predictors, we factored out frequency from *convergence* in the *True Recurrence* measure via division by *Combined Child and Adult Frequency* (F_a + F_c); and we factored out frequency from *convergence* in the *True Occurrence* measure by discarding occurrences of linguistic elements which fall within the area around the diagonal defined by $d = 2$ on the recurrence plots. This step ensures minimal relatedness of the two predictor variables, which in turn maximizes the independent effects of both. Still, one may find some amount of correlation between any two variables, for both random and unforeseen non-random reasons, and we use partial regression coefficients to control for this possibility.

In the first of two analyses, the coefficients are taken from linear regressions which model the growth of *Child Frequency*. If an increase in *True Adult Occurrence* results in a higher increase of *Child Frequency* than does an increase in *True Recurrence*, we can conclude that general frequency determines the child's usage more than *convergence*. On the other hand, if an increase in *True Recurrence* effects a larger increase in *Child Frequency* than would an increase in *True Adult Occurrence*, we may conclude that the impact of *convergence* on child usage of linguistic elements is greater than the impact of general frequency. In the second analysis, we perform the same kinds of calculations, except that this time we aim to predict *Adult Frequency*, and instead of *True Adult Occurrence* we use *True Child Occurrence* as a predictor; *True Recurrence* remains a predictor throughout.

If *True Recurrence* has any effect on usage, we would further like to distinguish between two kinds of recurrence: *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence*. They each track a different kind of *convergence*. *True Adult-Initiated Recurrence* tracks occurrences of elements which are used in *convergence* by the child following the adult's usage. That is, the child can here be seen to adapt to the adult. In contrast, *True Child-Initiated Recurrence* tracks occurrences of elements that are used in *convergence* by the adult after the child has used them. In other words, it tracks occurrences where the adult can be seen to adapt to the child. In so doing he or she may encourage future usage, by the child, of the element in question. In our analyses, if an increasing *True Adult-Initiated Recurrence* translates into more strongly increasing *Child / Adult Frequency* than does an increasing *True Child-Initiated Recurrence*, we may conclude that child-adaptation to the adult has a stronger impact on the usage of linguistic elements; if, however, an increase in *True Child-Initiated Recurrence* leads to a larger increase, it is probably the case that adult-adaptation to the child has the stronger impact.

We run the same analyses three times – once for content words, once for function words, and once for POS bigrams; and we do this for every one of the three corpora, so we end up with nine different analyses for predicting child usage, plus nine analyses for predicting adult usage. Requiring linguistic elements to occur both at least ten times in the adult's speech as well as at least ten times in the child's speech, we further exclude all elements from analysis that do not show a significant difference between *Recurrence* and *Random Recurrence* at $d = 2$, as established via a one-sided t-test, with $p \leq 0.05$. Sample sizes are given in Table 3.

Corpus	Content Words	Function Words	POS Bigrams
Abe	193 (552)	65 (172)	70 (226)
Adam	292 (553)	78 (168)	71 (195)
Sarah	254 (506)	53 (170)	50 (222)

Table 3: Sample sizes. The number to the left is the number of linguistic elements which show a significant difference between *Recurrence* and *Random Recurrence* at $d = 2$, and the number in parentheses is the total number of elements in the corpus.

Lastly, two additional notes about this study: (1) when using regression models to compare the impact of different predictors on *Child / Adult Frequency*, the models are based on transformed z-scores rather than the raw data. This is vital to the analysis because it standardizes the scales of values from different distributions, and in so doing makes coefficients based on different distributions comparable. For example, when using partial regression coefficients to compare the impact of *True Adult Occurrence* on *Child Frequency* to the impact of *True Recurrence* on *Child Frequency*, the distributions need to be one the same scale for the comparison the be meaningful. And (2) when referring to plots in order to illustrate patterns within the data, they are based on the Abe corpus. We are in a position to do things this way because the data pattern very similarly across the three corpora, and so plots based on the Abe corpus are representative of plots based on the other two corpora (but see the appendix for additional plots).

4 Analysis I – Child Usage

4.1 Frequency vs. Convergence

The goal of the first analysis was to compare the impact of *convergence* and general frequency in the adult's speech on child usage of linguistic elements, measured by their frequency in the child's speech. Thus, for each linguistic element, we calculate *Child Frequency*, *True Adult Occurrence* and *True Recurrence*. *True Adult Occurrence* measures frequency in the adult's speech, without considering elements used in *convergence*; *True Recurrence* tracks *convergence*; and *Child Frequency* is simply the frequency of an element in the child's speech.

We begin by modeling the data. Table 4 shows the slopes and, for completeness, R^2 values of the relevant multiple linear regression models. We see that across all corpora, the *True Adult Occurrence* coefficients for content words are generally much larger than the corresponding *True Recurrence* coefficients – meaning that *True Adult Occurrence*, independently of the other predictor, causes a larger increase in the average *Child Frequency* of content words than does an increase in *True Recurrence*.[14] For function words, *True Recurrence* leads by a small margin in the Abe corpus and trails by a similarly small margin in the Sarah corpus, while the *True Adult Occurrence* slope is twice as large as its *True Recurrence* counterpart in the Adam corpus. For POS bigrams, the two predictors take about equal values in the Sarah corpus, with *True Recurrence* being the stronger predictor in both the Abe and Adam corpus. On the whole, we find that *True Adult Occurrence* is the more important factor in determining *Child Frequency*, as its slope is the larger one in the majority of regression models.

[14]Note that here and in the later analyses, sometimes the smaller of the two partial regression coefficients is non-significant. In such cases, we still infer that the predictor associated with the larger coefficient has a stronger independent effect on the growth of the response variable, as it is the only predictor which independently contributes to the response variable.

Corpus	*Measure*	*True Adult Occurrence*	*True Recurrence*	R^2
Abe	Content Words	0.79 * * *	0.08	0.61
	Function Words	0.43 * * *	0.48 * * *	0.68
	POS Bigrams	0.20 *	0.72 * * *	0.78
Adam	Content Words	0.83 * * *	0.09 **	0.66
	Function Words	0.67 * * *	0.34 * * *	0.79
	POS Bigrams	0.20 *	0.71 * * *	0.76
Sarah	Content Words	0.85 * * *	0.18 * * *	0.66
	Function Words	0.57 * * *	0.50 * * *	0.77
	POS Bigrams	0.49 * * *	0.46 * * *	0.85

Table 4: Multiple linear regression models for the predictors *True Adult Occurrence* and *True Recurrence*, by measure and corpus. In the analysis, we are primarily interested in the slopes, but we include R^2 for completeness. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

Next, we inspect the plotted data. Figures 5 and 6 plot *Child Frequency* against *True Recurrence* (Figure 5) and *True Adult Occurrence* (Figure 6), for all three types of linguistic elements. For POS bigrams, the relationships between *True Recurrence* and *Child Frequency* as well as between *True Adult Occurrence* and *Child Frequency* appear linear, with the former being both more tightly correlated and forming a steeper line. We see a similar linear (but much less tightly correlated) relationship between *True Adult Occurrence* and *Child Frequency* for both content words and function words.

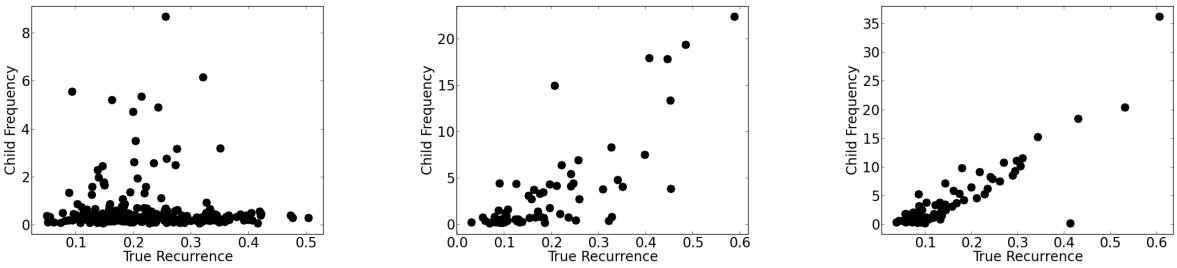

Figure 5: Scatterplots showing the relation between *True Recurrence* and *Child Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

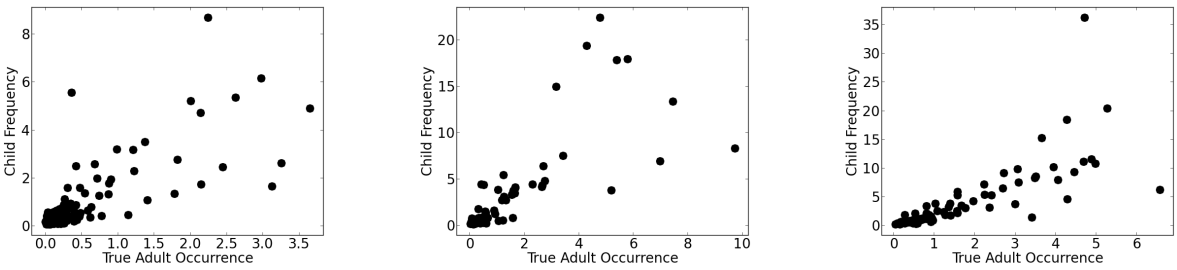

Figure 6: Scatterplots showing the relation between *True Adult Occurrence* and *Child Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

However, things are more complicated with Figures 5 (a) and (b), which plot *True Recurrence* against *Child Frequency*, for content and function words. With a lower tail at the bottom of

the plot, and a slanted, upwards-reaching part in the left region, Figure 5 (a) is reminiscent of a power distribution. We can discern a similar but less pronounced non-linear effect in Figure 5 (b), with a shorter lower tail and a more slanted upper part. The other plots, to all intents and purposes, appear more linear.[15]

Suspecting that high-frequency elements are concentrated in the upwards-reaching parts of the distributions shown in Figures 5 (a) and (b), with low-frequency words situated in the lower tails, we re-plot Figures 5 (a) – (c), but this time we color elements according to their *Combined Child and Adult Frequency*, i.e. $F_a + F_c$. Data points are colored black for maximal combined frequency, white for minimal combined frequency, and in shades of grey for intermediate values. Plotting the data in this way may give us an idea as to how frequency interacts with *True Recurrence.*

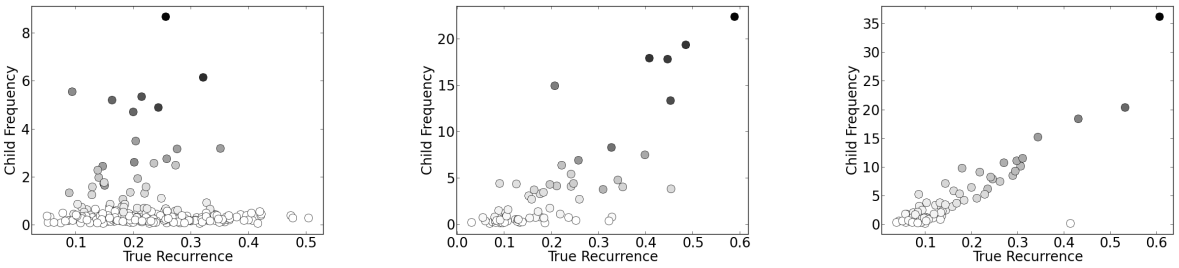

Figure 7: Scatterplots showing the relation between *True Recurrence* and *Child Frequency*.
A darker color indicates a higher *Combined Child and Adult Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

And indeed, the coloring reveals that elements in the lower tail of Figure 7 (a) appear to be infrequent (since they are colored more lightly), while data points in the left region appear to be more frequent (since they are colored more darkly). Looking closely at Figure 7 (b), we can discern a similar but less obviously apparent lower tail and a similar but less steep upwards-going slanted line, with data points seemingly split among the two in a similar fashion. We do not recognize such a pattern in Figure 7 (c).

Purely from inspecting the plotted data, we hypothesize that there exists an interaction between *Combined Child and Adult Frequency* and *True Recurrence*, in that the child's usage of high-frequency elements is more strongly affected by *True Recurrence* than is the usage of low-frequency elements. This is because if we were to imagine a partition of the data into a high-frequency and a low-frequency group – imagine all the darkly colored points in one group and the lightly colored data points in another group –, we would probably find that as we increase *True Recurrence*, the *Child Frequency* of linguistic elements in the high-frequency group would grow more than that of elements in the low-frequency group. That is: we would find that the impact of *True Recurrence* on *Child Frequency* is higher for elements in the high-frequency group.

To test this hypothesis, we wish to assign each linguistic element to one of the two groups, such that the high-frequency group contains the elements that occur most often and the low-frequency group contains the elements that occur least often. Furthermore,

[15]But note that almost all natural phenomena are likely to be non-linear to some extent. When we speak of a linear effect in some data set, we really mean that a linear function would be a good approximation; and when we speak of a non-linear effect, we really mean that a linear function would be a poor approximation. Thus, Figure 5 (a) appears less linear than the Figures 5 (b) – (c) and Figures 6 (a) – (c), and Figure 5 (b) appears less linear than Figure 5 (c) and Figures 6 (a) – (c); but since nature is awash in non-linear processes, the more linear-seeming figures are likely to be the products of non-linear processes themselves. In any event, the observations about non-linear effects in the data are only important in that they prompt further analytic probing. They should, therefore, not be overemphasized by the careful reader.

we would like our clustering procedure to allow for the scenario that the high-frequency group does not contain all and only the most frequent elements, and the low-frequency group does not contain all and only the least frequent elements. Rather, we would like the high-frequency group to include the most frequent elements as its nucleus, plus additional elements close to this nucleus, and similarly for the low-frequency group. After all, it could be the case that the data cannot be split evenly by frequency and e.g. the upper tail of Figure 7 (a) also contains a small number of relatively infrequent elements, in addition to a larger number of relatively frequent elements.

There are different ways to partition the data in a way that conforms to this requirement; and in choosing a method, we should keep in mind that we will eventually want to perform regression analyses on the two groups, with *True Recurrence* as the only predictor. This will enable us to compare the slopes of the regression lines for the high- and the low-frequency group, and to draw conclusions as to whether *True Recurrence* affects *Child Frequency* more strongly in one of the two clusters. But this can only be done if the two groups have a big-enough size, and if the regression models we wish to construct are statistically significant.

For this reason, our clustering heuristic derives many possible partitions of the data into a high- and a low-frequency group, and it picks a final partition such that (1) the high-frequency group has at least 30 members; (2) the high-frequency group has as its nucleus the most frequent elements, and the low-frequency group has as its nucleus the least frequent elements; (3) a significant ($p \leq 0.05$) regression model with *True Recurrence* as predictor can be fitted to elements in the high-frequency group; and (4) the partition maximizes the slope of the regression line on the high-frequency group. In case no partition exists that satisfies (1), we choose the one with the next-lowest number of elements (in the high-frequency group) that satisfies (2) – (4).

To derive the partitions, we define a variable n, which we initially set to 100. Then, presented with all the linguistic elements of a given type, we pick the elements on or above the nth percentile of the *Combined Child and Adult Frequency* distribution and fit a linear regression line to them. Call this the high-frequency line. A similar low-frequency line is obtained by fitting a regression model to the elements on and below the 5th percentile. Now, if an element is closer to the high-frequency line than to the low-frequency line, we assign it to the high-frequency group; otherwise, we assign it to the low-frequency group. We end up with a first partition, but we continue to derive 29 additional partitions by re-running the clustering procedure and decrementing n by one each time we do so.

That way, we derive a total of 30 different partitions where the nucleus of the high-frequency group is formed by elements alternatively falling on or above the 100th, 99th, 98th... 70th percentile of the *Combined Child and Adult Frequency* distribution. The nucleus of the low-frequency group, on the other hand, is always formed by elements on or below the 5th percentile. Out of the 30 partitions, we then pick the one that satisfies (1) – (4) above.

Figures 8 and 9 show the two groups we end up with, including regression lines fitted to elements in their groups with *True Recurrence* as predictor. Note that although we do not have an a priori reason to do so, we include POS bigrams in the analysis because appearances notwithstanding, there might still be a contrast between a high- and low-frequency group, similar to what we expect to find for content and function words.

The plots strongly suggest that our initial assumption is correct: as we increase *True Recurrence*, the *Child Frequency* of elements in the high-frequency group appears to increase more strongly than the *Child Frequency* of elements in the low-frequency group. Interestingly, we also find this contrast for POS bigrams, where the low-frequency group is less stretched-out along the x-axis and instead is more concentrated in the lower left. Group sizes are given in Table 5 below.

We obtain statistical evidence for the difference between the two groups by comparing

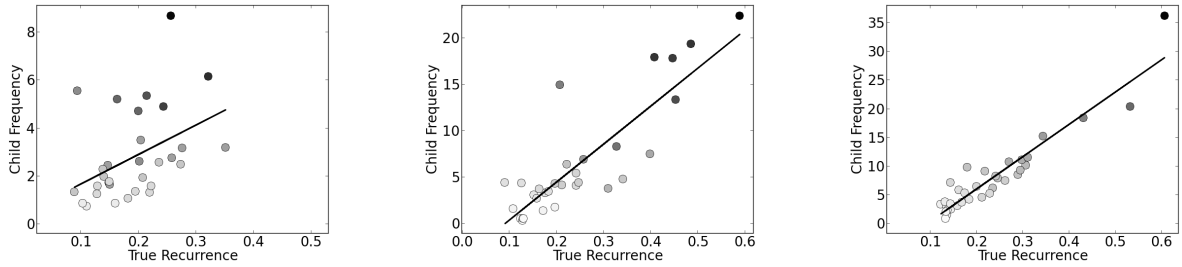

Figure 8: High-frequency groups. A darker color indicates a higher *Combined Child and Adult Frequency*. (a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

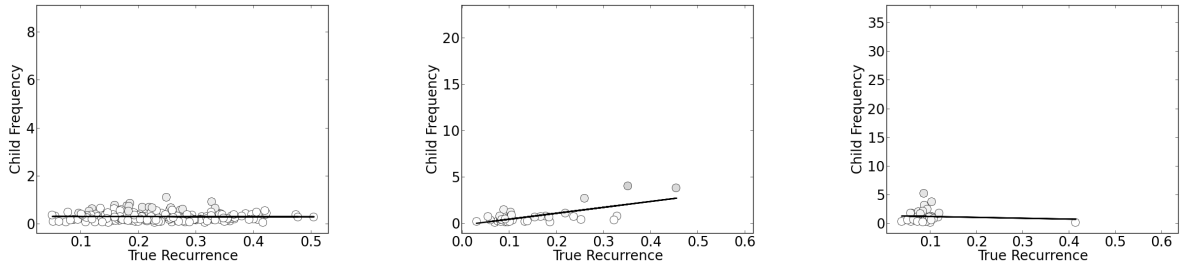

Figure 9: Low-frequency groups. A darker color indicates a higher *Combined Child and Adult Frequency*. (a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

Corpus	Content Words	Function Words	POS Bigrams
Abe	34; 159	31; 34	34; 36
Adam	29; 263	32; 46	33; 38
Sarah	35; 219	26; 27	30; 20

Table 5: Sizes of the low- and high-frequency groups, by corpus and element type. The first number is the size of the high-frequency group, and the second number is the size of the low-frequency group.

the slopes of the two regression lines. If (1) the regression model for the high-frequency group is significant, (2) its slope is larger, and (3) the difference between the two slopes is significant, we will have verified our hypothesis. Conditions (1) and (2) are easily met; and to test whether (3) holds, we create a dummy variable *group*, such that every element in the low-frequency group has value zero and every element in the high-frequency cluster has value one. Then, we run a multiple regression including the predictor *True Recurrence*, plus an interaction term of *True Recurrence* with *group*. If the regression coefficient for the interaction term is significant, the slope of the regression line for *True Recurrence* is significantly different in the two groups. Table 6 shows the differences of the regression coefficients (and R^2 values) for the high- and low-frequency groups. Because the differences are all positive as well as significant, we can conclude that the impact of *True Recurrence* on *Child Frequency* is indeed stronger for the high-frequency group. Further, it is noteworthy that the differences of R^2 values are also all positive; this means that we also obtained better fits for the high-frequency groups.

Corpus	Content Words	Function Words	POS Bigrams
Abe	0.46 (0.26) * * *	0.18 (0.29) * * *	0.94 (0.88) * * *
Adam	0.54 (0.26) * * *	0.59 (0.81) * * *	0.94 (0.88) * * *
Sarah	0.59 (0.53) * * *	0.35 (0.57) * * *	0.91 (0.82) * * *

Table 6: Differences of slopes and R^2 values (in parentheses) of linear regression models predicting *Child Frequency* for (1) the high-frequency group and (2) the low-frequency group, with *True Recurrence* as single predictor. A large positive difference signifies that the coefficient for the high-frequency group is larger. The stars refer to the significance level of the differences between the slopes. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

4.2 High-Frequency vs. Low-Frequency Elements

Since we have confirmed the hypothesis that there is a difference between the groups with respect to the impact of *True Recurrence* on *Child Frequency*, it would be interesting to see if within the high-frequency group, *True Recurrence* influences the growth of *Child Frequency* more strongly than *True Occurrence*. If this is the case, we can claim to have identified additional groups of linguistic elements whose usage is more strongly affected by *convergence* than by general frequency. Table 7 gives the details of multiple linear regressions obtained from the high-frequency groups.

Corpus	*Measure*	*True Adult Occurrence*	*True Recurrence*	R^2
Abe	Content Words	0.40 *	0.32 *	0.29
	Function Words	0.04	0.82 * * *	0.71
	POS Bigrams	-0.11	1.02 * * *	0.88
Adam	Content Words	0.45 **	0.36 *	0.42
	Function Words	0.07	0.90 * * *	0.93
	POS Bigrams	-0.11	1.02 * * *	0.88
Sarah	Content Words	0.47 * * *	0.59 * * *	0.75
	Function Words	0.54 * * *	0.52 * * *	0.80
	POS Bigrams	0.34	0.60 **	0.84

Table 7: Multiple linear regression models for the predictors *True Adult Occurrence* and *True Recurrence*, by measure and corpus. Models are based on the high-frequency groups.
*** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

For Abe, as the slope is larger for the regression models based on *True Recurrence*, the *Child Frequency* of function words and POS bigrams is more strongly impacted by *True Recurrence* than by *True Adult Occurrence*. The trend is reversed with content words, whose usage is (still) more strongly influenced by *True Adult Occurrence*. The situation is almost identical for Adam, with minor differences in the actual numbers, but with the same general pattern. And for Sarah, *True Recurrence* is the stronger influence with content words and POS bigrams, while the two predictors exert a similar influence on her usage of function words. With the usage of function words in the Adam corpus as well as POS bigrams and content words in the Sarah corpus now more strongly influenced by *True Recurrence*, *True Recurrence* has become the strongest general predictor. Its coefficient is the largest in more than half of the models – a change from the state of affairs when we considered high- and low-frequency elements together.

Finally, the clustering method we used only stipulated that the high-frequency group should have as its nucleus the elements with the highest combined frequency, and similarly for the low-frequency group. But is it the case that combined frequency completely determines group membership? To check this, we plot the Zipfian *Combined Child and Adult Frequency* distributions of (1) all data points, (2) the high-frequency group, and (3) the low-frequency group. The resultant plots look relatively similar across corpora and measures. Figure 10 below, based on function words from the Abe corpus, serves as an illustration of the typical pattern: the groups have fuzzy boundaries, but elements are rather evenly split by combined frequency. This lends support to the conclusion that the *Child Frequency* of high-frequency elements is, at least for certain types of linguistic elements and with some idiosyncratic variation across children, more strongly affected by *True Recurrence* than by *True Adult Occurrence*.

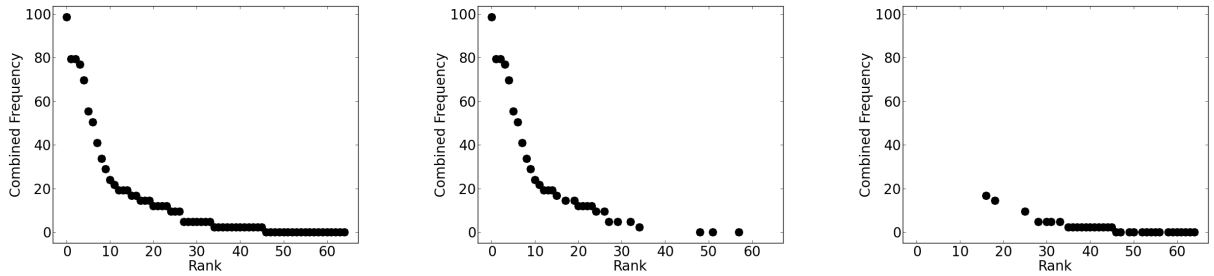

Figure 10: Zipfian *Combined Child and Adult Frequency* distributions. The 'steps' in the lower tail are due to average *Combined Child and Adult Frequency* values being rounded.
(a) Left: All Data Points – (b) Middle: High-Frequency Group – (c) Right: Low-Frequency Group

4.3 Adult-Initiated vs. Child-Initiated Convergence

Having established that *True Recurrence* has a strong impact on *Child Frequency*, that this impact is stronger for linguistic elements in the high-frequency group, and that for some groups of elements the impact is actually stronger than that of *True Adult Occurrence*, a natural follow-up question is how *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence* fit into the picture. Hence, the goal of this section is to explore the general pattern with respect to these two measures and to investigate whether there is a difference between them.

Recall that *True Adult-Initiated Recurrence* corresponds to the likelihood of an element being used in *convergence* by the child after the adult has used it, whereas *True Child-Initiated Recurrence* corresponds to the likelihood of an element being used in *convergence* by the adult after the child has used it. Thus, the distinction between between the two can be seen to break down *True Recurrence* into a part where the adult initiates – or leads – usage in *convergence* and a part where the child is leading. Establishing the effect of these two measures on *Child Frequency* will allow us to see whether child-adaptation to the adult (*True Adult-Initiated Recurrence*) is more important in determining the child's usage of linguistic elements, or whether adult-adaptation to the child (*True Child-Initiated Recurrence*) is the more important determinant. In order to get an impression of the overall pattern, we again plot the data. Figure 11 plots *Child Frequency* against *True Child-Initiated Recurrence*, and Figure 12 plots *Child Frequency* against *True Adult-Initiated Recurrence*. Elements are colored according to their *Combined Child and Adult Frequency*.

Two observations can be made right away: (1) the pattern for *True Child-Initiated Recurrence* seems very similar to the pattern for *True Adult-Initiated Recurrence*, and (2) both patterns look very similar to what we found for *True Recurrence*. This should, however, not

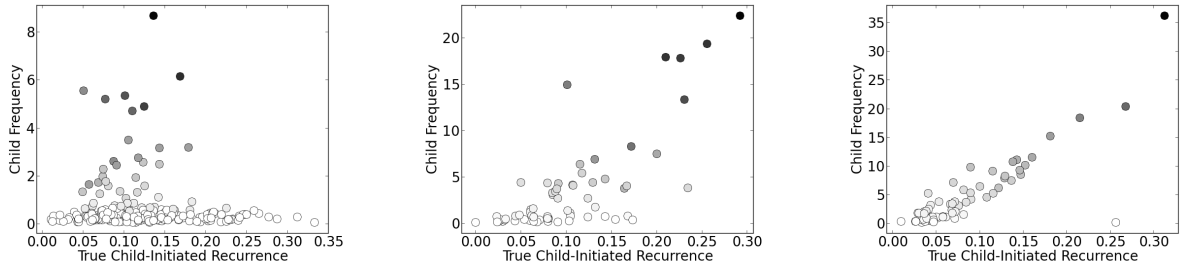

Figure 11: Scatterplots showing the relation between *True Child-Initiated Recurrence* and *Child Frequency*. A darker color indicates a higher *Combined Child and Adult Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

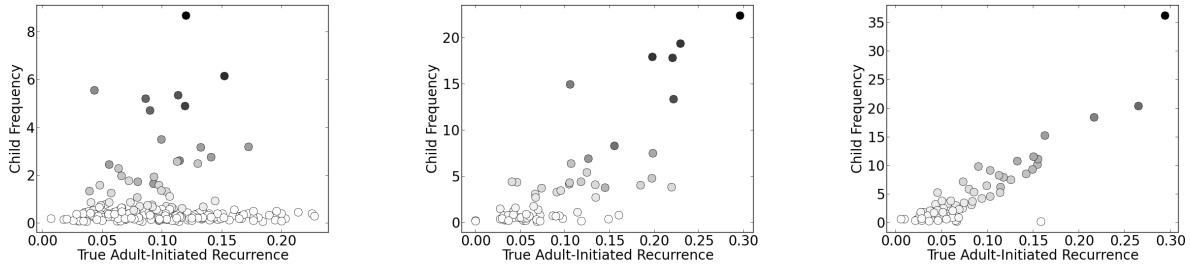

Figure 12: Scatterplots showing the relation between *True Adult-Initiated Recurrence* and *Child Frequency*. A darker color indicates a higher *Combined Child and Adult Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

incline us to expect only minor differences between the two. While the overall impact of the two predictors might be similar, they may each make quite different independent contributions to the growth of *Child Frequency*. This suspicion is confirmed by the regression models summarized in Table 8, which predict *Child Frequency* from *True Adult-Initiated Recurrence*, *True Child-Initiated Recurrence*, and *True Adult Occurrence*.

Corpus	*Measure*	*True Adult-Init. Recurrence*	*True Child-Init. Recurrence*	*True Adult Occ.*	R^2
Abe	Content Words	0.03	0.05	0.79 * * *	0.61
	Function Words	0.27	0.22	0.43 * * *	0.68
	POS Bigrams	0.90 * * *	-0.10	0.11	0.80
Adam	Content Words	0.01	0.09 *	0.83 * * *	0.66
	Function Words	0.34 * * *	0.06	0.60 * * *	0.80
	POS Bigrams	0.50 * * *	0.24	0.18 *	0.76
Sarah	Content Words	0.08	0.11 *	0.85 * * *	0.66
	Function Words	0.35 * * *	0.19 *	0.55 * * *	0.77
	POS Bigrams	0.20	0.28 *	0.48 * * *	0.85

Table 8: Multiple linear regression models for the predictors *True Adult-Initiated Recurrence*, *True Child-Initiated Recurrence* and *True Adult Occurrence*, by measure and corpus. Models are based on all data points. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

In essence, these models are carried over from section 4.1, except that *True Recurrence* is replaced by *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence*, the sum of whose coefficients is very close to the *True Recurrence* slopes from Table 4; the R^2 values

are also either identical or very close to the values reported there. By comparing the coefficients for *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence*, we may ascertain which of these predictors has the stronger independent impact on *Child Frequency*. Due to the negligibly small effect size, we ignore results for content words and focus first on Abe and Adam. Disregarding cases from the Abe corpus where both *Recurrence* predictors are non-significant (in which case neither predictor makes an independent contribution to the growth of the response variable), we notice right away that for Abe and Adam, *True Adult-Initiated Recurrence* is the stronger of the two predictors. In the Sarah corpus, we also find this to be the case for function words, although the pattern is reversed for POS bigrams. Taken together, these results indicate that child-adaptation to the adult (*True Adult-Initiated Recurrence*) may be the more important factor in determining *Child Frequency*.

If true, the same pattern should emerge also for high-frequency elements. Since the general pattern for *True Recurrence* is so similar to the patterns for *True Child-Initiated Recurrence* and *True Adult-Initiated Recurrence*, we restrict analysis to the high- and low-frequency groups that were established in the previous analysis, where we plotted *True Recurrence* against *Child Frequency*. This simplifies the analysis in that rather than partitioning the data three times – once for *True Recurrence*, once for *True Child-Initiated Recurrence*, and once for *True Adult-Initiated Recurrence* –, we can simply continue working with the partition for *True Recurrence*.[16]

Having established that here as well, the slopes of the regression lines with *True Child-Initiated Recurrence* and *True Adult-Initiated Recurrence* as predictors differ significantly between the high- and low-frequency groups, we can proceed to compare partial regression coefficients for the two predictors on the high-frequency group in Table 9.

Corpus	*Measure*	*True Adult-Init. Recurrence*	*True Child-Init. Recurrence*	*True Adult Occ.*	R^2
Abe	Content Words	-0.07	0.39	0.47	0.28
	Function Words	0.31	0.52	0.04	0.70
	POS Bigrams	0.26	0.75	-0.01	0.88
Adam	Content Words	0.67	0.29	0.49 **	0.41
	Function Words	0.64 **	0.26	0.78	0.93
	POS Bigrams	0.81 ***	0.25	-0.14	0.88
Sarah	Content Words	0.42 *	0.20 *	0.45 ***	0.75
	Function Words	0.42 **	0.16	0.51 ***	0.81
	POS Bigrams	0.28	0.30	0.34	0.83

Table 9: Multiple linear regression models for the predictors *True Adult-Initiated Recurrence*, *True Child-Initiated Recurrence* and *True Adult Occurrence*, by measure and corpus. Models are based on the high-frequency groups. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

High p values forbid meaningful interpretation of most of the results; still, where they are statistically significant, the coefficients for *True Adult-Initiated Recurrence* are larger than their *True Child-Initiated Recurrence* counterparts. While perhaps not the strongest evidence, this at least does not diminish the plausibility of child-adaptation to the adult (*True Adult-Initiated Recurrence*) being the more important factor in shaping the child's usage of linguistic elements.

[16] Actually deriving the three partitions and comparing groups confirms this approach: the high-frequency groups for the three measures usually only differ by two to five elements.

5 Analysis II – Adult Usage

5.1 Frequency vs. Convergence

In this second analysis, we perform the same computations as in analysis I – the difference being that instead of the child's speech, the focus is on the *adult's* speech. Accordingly, the response variable changes from *Child Frequency* to *Adult Frequency*, i.e. the average frequency with which a given linguistic element occurs in the adult's speech across all dialogue transcripts in a corpus. We keep *True Recurrence* as a predictor, and we exchange *True Adult Occurrence* for *True Child Occurrence*. But except for these changes, the analysis remains exactly the same as before.

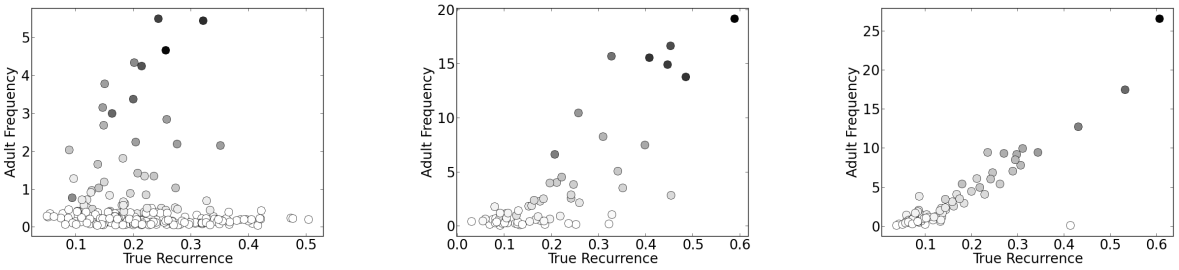

Figure 13: Scatterplots showing the relation between *True Recurrence* and *Adult Frequency*. A darker color indicates a higher *Combined Child and Adult Frequency*
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

The general pattern we find is very similar to the one uncovered in analysis I. Here as there, elements with a higher *Combined Child and Adult Frequency* are concentrated in a slanted, upwards-reaching group of data points in the left quadrant of the scatterplot for content words. This is shown in Figure 13, which plots *True Recurrence* against *Adult Frequency*. As in analysis I, elements are colored according to their *Combined Child and Adult Frequency*.

Corpus	*Measure*	*True Child Occurrence*	*True Recurrence*	R^2
Abe	Content Words	0.80 * * *	0.01	0.63
	Function Words	0.43 * * *	0.54 * * *	0.73
	POS Bigrams	0.30 * * *	0.65 * * *	0.83
Adam	Content Words	0.86 * * *	0.03	0.74
	Function Words	0.73 * * *	0.24 * * *	0.80
	POS Bigrams	0.23 **	0.71 * * *	0.79
Sarah	Content Words	0.82 * * *	-0.11 * * *	0.72
	Function Words	0.82 * * *	-0.01	0.65
	POS Bigrams	0.24 **	0.71 * * *	0.85

Table 10: Multiple linear regression models for the predictors *True Child Occurrence* and *True Recurrence*, by measure and corpus. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

Table 10 contains the coefficients and R^2 values of the relevant multiple linear regressions. Without separating the data into a high-frequency and a low-frequency group, *True Child Occurrence* is the better predictor as far as content words are concerned – the same pattern we discovered in analysis I. Also similar to what was found analysis I, *True Recurrence* results

in higher slopes for POS bigrams. The *Adult Frequency* of function words is much more strongly affected by *True Child Occurrence* in both the Adam and the Sarah corpus, while *True Recurrence* leads by a small margin in the Abe corpus. And although results from analysis I are obviously not exactly replicated, *True Adult Occurrence* is associated with the larger slope in the majority of regression models – just as *True Child Occurrence* in analysis I.

Corpus	Content Words	Function Words	POS Bigrams
Abe	34; 159	33; 32	35; 35
Adam	51; 241	32; 46	32; 39
Sarah	14; 240	30; 23	34; 16

Table 11: Sizes of the low- and high-frequency groups, by corpus and element type. The first number is the size of the high-frequency group, and the second number is the size of the low-frequency group.

At this stage, we again separate the data into a high-frequency group and a low-frequency group, where the most frequent elements form the core of the first and the least frequent elements form the core of the second group (group sizes are given in Table 11). With *True Recurrence* as the only predictor, we fit regression models to each of the two clusters we obtain per corpus and element type, ascertain whether the slopes are significantly different ($p \leq 0.05$) via the procedure described in section 4.1, and take the differences of the regression coefficients, which are given in Table 12.

Corpus	Content Words	Function Words	POS Bigrams
Abe	0.70 (0.48) * * *	0.28 (0.45) * * *	0.96 (0.93) * * *
Adam	0.64 (0.61) * * *	0.64 (0.81) * * *	0.96 (0.92) * * *
Sarah	0.55 (0.20) * * *	0.62 (0.36) *	0.93 (0.82) * * *

Table 12: Differences of slopes and R^2 values (in parentheses) of linear regression models predicting *Adult Frequency* for (1) the high-frequency group and (2) the low-frequency group, with *True Recurrence* as single predictor. A large positive difference signifies that the coefficient for the high-frequency group is larger. The stars refer to the significance level of the differences between the slopes. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

It should not come as a surprise that as with *Child Frequency* in the foregoing analysis I, the impact of *True Recurrence* on *Adult Frequency* is larger for elements in the high-frequency group than for elements in the low-frequency group. The differences of the slopes for regression models fitted on the two groups, with *True Recurrence* as the sole predictor, are all positive – and statistically significant ($p \leq 0.05$). And except for POS bigrams, where the results are about the same, the differences are even larger than what we found in analysis I. Beyond these differences, the pattern is again remarkably similar to the one uncovered in analysis I.

5.2 High-Frequency vs. Low-Frequency Elements

Given that the differences between the *True Recurrence* slopes for the two groups are again all positive, it is possible that within the high-frequency group, *True Recurrence* has a larger impact on *Adult Frequency* than *True Child Occurrence*. Table 13 largely verifies this suspicion. In the Abe corpus, *True Recurrence* has a larger impact on the growth of the *Adult Frequency* of all three element types. Compared to the results from analysis I, the overall pattern for

Abe has changed in that *True Recurrence* now also has a stronger impact on the *Adult Frequency* of content words. For Adam, the effect of the two predictors is about the same for content words (a minor change from analysis I, where *True Adult Occurrence* had the stronger impact), while it is markedly stronger for both POS bigrams and function words (as in analysis I). In the Sarah corpus, lastly, the *Adult Frequency* of POS bigrams in the high-frequency group is still more strongly impacted by *True Recurrence*, while *True Adult Occurrence* now has the stronger impact on the frequency of function words, where previously we diagnosed an equal influence. All in all – and just as in analysis I – *True Recurrence* is the dominant predictor for elements in the high-frequency group, as its slope is the larger one in more than half of the regression models.

Corpus	*Measure*	*True Child Occurrence*	*True Recurrence*	R^2
Abe	Content Words	0.36 **	0.52 * * *	0.57
	Function Words	0.11	0.84 * * *	0.83
	POS Bigrams	0.05	0.92 * * *	0.92
Adam	Content Words	0.45 * * *	0.47 * * *	0.73
	Function Words	0.21	0.74 * * *	0.89
	POS Bigrams	-0.08	1.03 * * *	0.92
Sarah	Content Words	0.13	0.43	0.18
	Function Words	0.95 * * *	-0.04	0.82
	POS Bigrams	0.16	0.81 * * *	0.87

Table 13: Multiple linear regression models for the predictors *True Child Occurrence* and *True Recurrence*, by measure and corpus. Models are based on the high-frequency groups.
*** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

Finally, as in section 4.2, we plot the Zipfian *Combined Child and Adult Frequency* distributions of (1) all data points and (2) the high- and low-frequency groups. As we might have expected, results do not differ much across corpora and element types, and so we contend ourselves with providing plots based on content words from the Abe corpus. Figure 14, which based on function words in the Abe corpus, clearly shows that while it does not completely determine cluster-membership, linguistic elements are still rather evenly split by *Combined Child and Adult Frequency*.

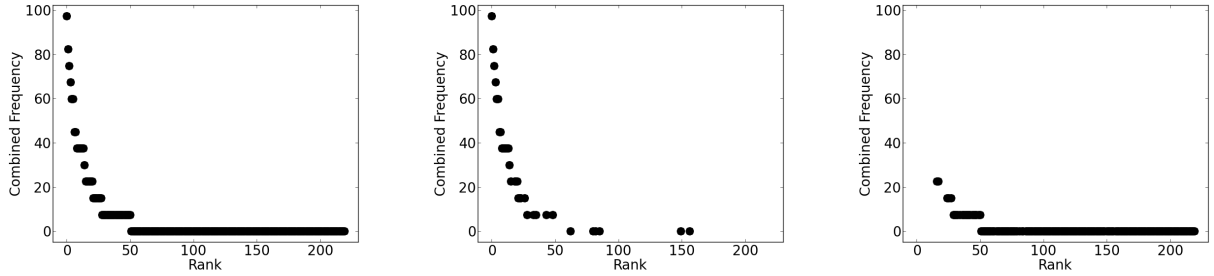

Figure 14: Zipfian *Combined Child and Adult Frequency* distributions. The 'steps' in the lower tail are due to average *Combined Child and Adult Frequency* values being rounded.
(a) Left: All Data Points – (b) Middle: High-Frequency Group – (c) Right: Low-Frequency Group

5.3 Adult-Initiated vs. Child-Initiated Convergence

As the last step in the analysis of how *True Recurrence* and *True Child Occurrence* affect *Adult Frequency*, we break down *True Recurrence* into *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence*. Just as in section 4.3, the goal is to compare how these two measures impact *Adult Frequency*.

Figures 15 and 16 assure us in our expectation that we will not discover marked differences to the pattern identified in section 4.3. There, we observed that (1) the situations with respect to *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence* are quite similar and that (2) the patterns for both these predictors are themselves very similar to the pattern we find with *True Recurrence*. Looking at Figures 15 and 16, which plot the two predictors against *Adult Frequency*, this appears to be the case here as well.

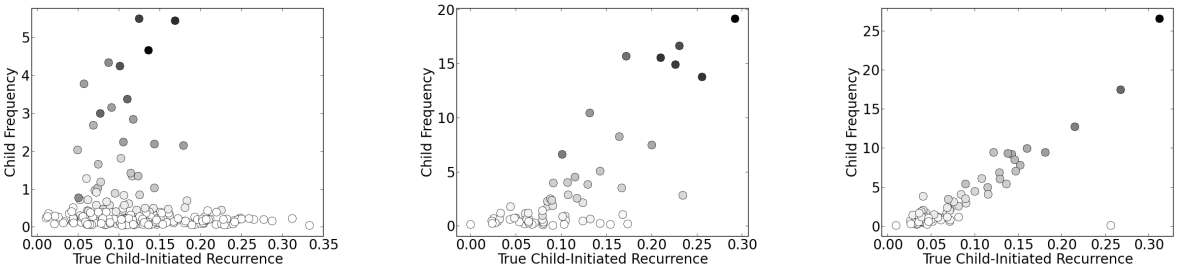

Figure 15: Scatterplots showing the relation between *True Child-Initiated Recurrence* and *Adult Frequency*. A darker color indicates a higher *Combined Child and Adult Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

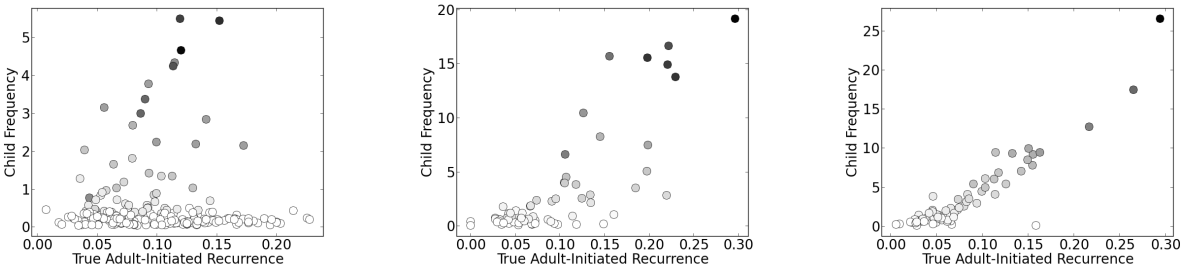

Figure 16: Scatterplots showing the relation between *True Adult-Initiated Recurrence* and *Adult Frequency*. A darker color indicates a higher *Combined Child and Adult Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

As before, we perform regression analyses with the predictors *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence*; the third predictor, this time, is *True Child Occurrence*; and our response variable is *Adult Frequency*. A look at Table 14 confirms that the sum of the two *Recurrence* slopes is very close to the *True Recurrence* slopes from the models that included *True Recurrence* and *True Child Occurrence* as predictors (Table 10). That is, we essentially replace *True Recurrence* with *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence*, and we can again determine which of the two predictors is the more impactful variable by directly comparing their values.

Naturally, we ignore non-significant results; and because of the small effect size, we also ignore results for content words. With this in mind, we find that *True Adult-Initiated Recurrence* appears to be the more important variable in determining the growth of *Adult Frequency*. Its values are larger for function words and POS bigrams in the Abe and Adam

Corpus	*Measure*	*True Adult-Init. Recurrence*	*True Child-Init. Recurrence*	*True Child Occ.*	R^2
Abe	Content Words	0.08	-0.05	0.79 ***	0.63
	Function Words	0.32 *	0.23	0.42 ***	0.77
	POS Bigrams	0.87 ***	-0.15	0.23 **	0.85
Adam	Content Words	0.08 *	-0.04	0.86 ***	0.74
	Function Words	0.26 ***	0.04	0.67 ***	0.80
	POS Bigrams	0.49 ***	0.24 *	0.21 **	0.79
Sarah	Content Words	≈ 0	-0.11 **	0.82 ***	0.72
	Function Words	0.06	-0.06	0.81 ***	0.64
	POS Bigrams	0.30 **	0.45 ***	0.25 **	0.85

Table 14: Multiple linear regression models for the predictors *True Adult-Initiated Recurrence*, *True Child-Initiated Recurrence* and *True Adult Occurrence*, by measure and corpus. Models are based on all data points. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

corpus, whereas the coefficient for *True Child-Initiated Recurrence* is only larger for POS bigrams in the Sarah corpus. We found in section 4.3 that *True Adult-Initiated Recurrence* is the more important determinant of *Child Frequency*, and it would appear that it is also the more important factor in determining the frequency of linguistic elements in the adult's speech.

Is this pattern substantiated when we restrict analysis to the high-frequency groups? For simplicity, and because the high- and low-frequency groups overlap to a large extent, we continue to work with the groupings derived in section 5.1. As expected, the high-frequency slopes for *True Adult-Initiated Recurrence* and *True Child-Initiated Recurrence* are significantly different ($p \leq 0.05$) from their low-frequency counterparts across all corpora and measures, with the exception of function words in the Sarah corpus (although $p \leq 0.1$). Given this, we proceed to comparing the high-frequency slopes. Table 15 below shows the partial regression coefficients and R^2 values of the relevant models.

Corpus	*Measure*	*True Adult-Init. Recurrence*	*True Child-Init. Recurrence*	*True Child Occ.*	R^2
Abe	Content Words	0.53 *	≈ 0	0.36 **	0.58
	Function Words	≈ 0	0.86 **	0.10	0.85
	POS Bigrams	0.39	0.58	0.05	0.92
Adam	Content Words	0.37 **	0.13	0.46 ***	0.73
	Function Words	0.49 *	0.26	0.21	0.88
	POS Bigrams	0.73 ***	0.33 *	-0.11	0.93
Sarah	Content Words	0.36	0.10	0.11	0.10
	Function Words	≈ 0	-0.04	0.94	***
	POS Bigrams	0.21	0.60 ***	0.16	0.87

Table 15: Multiple linear regression models for the predictors *True Adult-Initiated Recurrence*, *True Child-Initiated Recurrence* and *True Adult Occurrence*, by measure and corpus. Models are based on the high-frequency groups. *** : $p \leq 0.001$, ** : $p \leq 0.01$, * : $p \leq 0.05$

Overall, the coefficients for *True Adult-Initiated Recurrence* are still larger. This is the case for content words in the Abe corpus and all three types of linguistic elements in the Adam corpus. *True Child-Initiated Recurrence* is still the stronger predictor for POS bigrams in the

Sarah corpus; and reversing the trend from Table 14, it now also is the stronger predictor for function words in the Abe corpus. Nevertheless, *True Adult-Initiated Recurrence* still weights heavier in the majority of (statistically significant) cases.

6 Discussion

6.1 General Observations and Inferences

We can make four general observations based on the results produced by the two analyses. (1) First, the usage of some types of linguistic elements is more strongly influenced by their frequency in the other interlocutor's speech, whereas the usage of other element types is more strongly influenced by how likely to they are to be used in *convergence*. (2) Second, and discounting some individual differences, the patterns for the three corpora are all very similar. (3) Third, the way in which the child's language use is influenced by *convergence* and frequency in the other interlocutor's speech closely resembles the way in which the adult's language use is influenced by these two factors. The resemblance is so strong, in fact, that in discussing the results, we often do not need to differentiate between child and adult usage at all, referring instead simply to *usage*. (4) Fourth, the importance of *Adult-Initiated Recurrence* relative to *Child-Initiated Recurrence* suggests that child-adaptation to the adult is more important than adult-adaptation to the child, both in determining the adult's and in determining the child's language use. We address the four points in turn.

1. The influence of frequency and convergence on different groups of linguistic elements

Without splitting the data into a high- and low-frequency group, the usage of content words is most strongly affected by their frequency in the other interlocutor's speech; this is followed by function words, whose usage is still somewhat more strongly affected by frequency; the usage of POS bigrams, finally, is actually more strongly influenced by *convergence*. This suggests that the usage of elements which are more semantic in nature (content words) is more strongly impacted by frequency, while the usage of more syntactic elements (POS bigrams) is more heavily influenced by *convergence*. Function words, which are closer to the semantic end of the spectrum than POS bigrams yet also are clearly not entirely semantically-natured, are neither as strongly impacted by frequency as are content words, nor are they as strongly affected by *convergence* as are POS bigrams. We will perform some additional analyses in the following section 6.2, in order to explore more closely the influence of frequency and *convergence* on the usage of the three different element types.

For high-frequency elements, the impact of *convergence* is increased, even though content words are still least affected by it, followed by function words, which in turn are followed by POS bigrams. Overall, high-frequency content words are about evenly strongly influenced by frequency and *convergence*; for high-frequency function words, the pattern changes in that they are now more strongly influenced by *convergence*; and the usage of high-frequency POS bigrams is so strongly affected by *convergence* that frequency either only plays a minor part in shaping their usage (Sarah corpus), or its influence is virtually non-existent (Abe and Adam corpus). This dominance of *convergence* in determining the usage of high-frequency elements may well mean that compared to frequency, the majority of both interlocutors' dialogue contributions are shaped through *convergence*. After all, the most frequent linguistic elements account for a disproportionately large part of the actual language produced.

2. Similarities across corpora

For all three corpora, the way in which frequency and *convergence* affect the frequency of linguistic elements in the other interlocutor's speech is very similar. It is striking how sim-

ilarly the data pattern on the scatterplots for the different corpora in the appendix, and the regression coefficients from the two analyses also usually take similar values. Generally, the patterns identified above hold for all three corpora: across corpora, frequency in the other interlocutor's speech is the more important determinant of the usage of most linguistic elements, although most high-frequency elements are actually more strongly influenced by *convergence*. In addition, POS bigrams are subject to the strongest influence of *convergence*, followed by function words, which in turn are followed by content words.

Some exceptions to this are found in the Sarah corpus, where compared to the other two corpora, frequency appears to play a more important role. For example, *True Recurrence* and *True Adult Occurrence* have about an equally strong impact on the frequency with which Sarah uses POS bigrams, whereas *True Recurrence* is by far the more important determinant of both *Child Frequency* and *Adult Frequency* in the corresponding analyses on the Abe and Adam corpora. Similarly, within the Sarah corpus, *True Recurrence* appears to have no independent impact on the *Adult Frequency* of function words, even though the corresponding analyses from the other two corpora picked up on an appreciable impact.

But aside from this, it is remarkable how similar the patterns are across the three corpora, with the differences between corresponding partial regression coefficients often being no higher than 0.1. Of course, given that we investigated only three child-caregiver dyads, we cannot conclude that we have unearthed universal patterns. Nevertheless, given that the similarities outweigh the differences, it is not implausible that also for most other child-caregiver dyads, we would find that both frequency and *convergence* play a part in shaping the other interlocutor's speech, with more frequent elements being more strongly influenced by *convergence*, and thus perhaps with most actual dialogue contributions being affected more strongly by *convergence* than by frequency. The exceptions found in the Sarah corpus may then incline us to also expect some idiosyncratic variation in how dyads match each other's language use.

3. The symmetry in how caregiver and child speech are affected

When comparing how frequency and *convergence* affect the frequency of elements in the adult's speech to how the two factors affect the frequency of elements in the child's speech, we are again struck by how similar the patterns are. The actual patterns, which we described above, remain unchanged – whether we consider the effect on adult language or the effect on child language. That is to say, not only are corresponding partial regression coefficients very similar across corpora, they *also* do not usually differ by more than 0.1 as we go from considering the effect on the child's speech to considering the effect on the adult's speech.

The single exception to this are again the data from the Sarah corpus. There, as far as e.g. function words are concerned, *True Child Occurrence* is much more important for determining *Adult Frequency* than is *True Adult Occurrence* for determining *Child Frequency*. In other words, frequency in the other interlocutor's speech is more important for determining adult usage of function words than for determining child usage of function words. More generally, within the Sarah corpus, the way in which the child's language use influences the adult's language use is less closely mirrored in how the adult's language use affects the child's language use. In the other corpora, the ways in which child and adult language are influenced by the speech of the other interlocutor resemble one another more closely.

The similarities in how the two interlocutors are influenced by each other suggest that the same mechanisms may be at work when the child's language use is affected by the adult's and when the adult's language use is affected by the child's. Part of the speech of both child and adult may thus be the product of a general adaptation process which operates both globally, through frequency, and locally, through *convergence*. In the end, the child's speech may be a reflection of the adult's speech, which given the child's linguistic limitations will

obviously not always be very close to the adult's actual language use. In much the same way, child-directed speech (CDS) may in part simply be a reflection of the child's speech, brought about by the very same mechanisms that shape the child's speech.

4. The importance of child-adaptation to the caregiver

When comparing the relative importance of *Child-Initiated Recurrence* to that of *Adult-Initiated Recurrence*, the latter clearly wins out over the former: in the majority of regression analyses, the coefficient for *Adult-Initiated Recurrence* is larger than the coefficient for *Child-Initiated Recurrence*. Interestingly, this is the case both when looking at the effect on adult language as well as when considering the effect on child language.

We may infer from this that child-adaptation to the adult is the more important determinant in shaping both the child's and the adult's speech. For example, if the adult uses a specific word and the child re-uses this word in the next turn, this is more likely to lead to future usage of the same word by the child than the child using the word and the adult re-using it in the next turn. In addition, the child's re-using of a given word is *also* more likely to lead to an increased usage of the element by the adult. That is, if the child re-uses a word shortly after the adult has used it, the adult will be more likely to use this word in future turns; and he will be more likely to do so, by virtue of the child adapting to him, than if the child had used the word in question and the adult had re-used it shortly thereafter. Child-adaptation to the adult, it seems, is more important than adult-adaptation to the child in determining the usage of linguistic elements by both the child *and* the adult.

This suggests that the two interlocutors are sensitive to each other's language use in different ways. The caregiver may be more sensitive to how the child responds to his utterances – i.e., he actually adapts to how the child adapts to *him*. The child, on the other hand, may more directly adapt to the adult's utterances, without being as sensitive to how the adult responds to her. Perhaps, through immediately re-using them after the adult, the child re-enforces linguistic elements in her developing language system, ultimately using them more and more often. The caregiver, on the other hand, might tailor future utterances to contain more of those elements to which the child responded favorably by immediately re-using them.

6.2 Some Post-Hoc Explorations

An issue we have not pursued so far is why the data pattern as they do with respect to *True Recurrence*. As we saw in sections 4 and 5, content words, function words, and POS bigrams pattern differently. Reproduced from section 4.1, Figure 17 below is based on the Abe corpus and plots *True Recurrence* against *Child Frequency*. As we go from content words to function words to POS bigrams, the distributions of linguistic elements becomes more linear, with data points centering on a straight line, going from the lower left corner to the upper right corner of the plot. This line is most pronounced for POS bigrams (Figure 17 (c)), still somewhat visible for function words (Figure 17 (b)), and least obvious (if at all present) for content words (Figure 17 (a)). In effect, the straight line from Figure 17 (c) is beginning to split into two separate lines in Figure 17 (b), which is only fully apparent in Figure 17 (a), where the partition into high- and low-frequency elements is clearest, with a flat line of low-frequency elements at the bottom of the figure and an upwards-going slanted distribution of more frequent elements in the left. In this subsection, we explore why the data pattern so differently for the three element types, as well as why some elements – even those within the flat low-frequency lines at the bottom of Figures 17 (a) and (b) – have a higher chance of recurring than others. The two questions are addressed in turn.

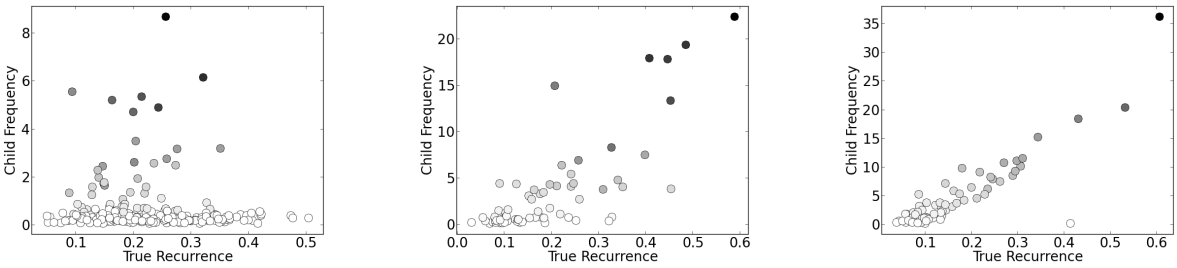

Figure 17: Scatterplots showing the relation between *True Recurrence* and *Child Frequency*.
A darker color indicates a higher *Combined Child and Adult Frequency*.
(a) Left: Content Words – (b) Middle: Function Words – (c) Right: POS Bigrams

1. Why do the data pattern differently for the three element types?

Frequency is likely responsible for at least some of the differences in how the data pattern, since we should expect there to be many more low-frequency content words than function words, and possibly more low-frequency function words than POS bigrams. Table 16 below shows the average frequency of the different element types per dialogue, largely confirming our expectations: across all corpora, content words are the least frequent; and in the Abe and Adam corpus, function words are the next most-frequent, followed by POS bigrams as the most frequent element type. In the Sarah corpus, function words are slightly more frequent than POS bigrams, although the data pattern similarly to what is shown in Figure 17 (c) above (cf. the appendix for plots based on the Adam and Sarah corpora).

Corpus	Content Words	Function Words	POS Bigrams
Abe	1.18 (sd: 1.94)	6.51 (sd: 9.45)	7.9 (sd: 10.10)
Adam	4.02 (sd: 6.17)	14.74 (sd: 27.61)	25.93 (sd: 39.05)
Sarah	1.62 (sd: 2.74)	8.08 (sd: 12.73)	7.24 (sd: 9.08)

Table 16: Average *Combined Child and Adult Frequency* of linguistic elements per dialogue transcript, by element type and corpus. For example, without separating adult and child speech, any given function word occurs on average 4.02 time per dialogue transcript in the Adam corpus. The larger values for the Adam corpus result from there being a smaller number of dialogue transcripts in that corpus, which at the same time are longer than in the other two corpora.

Since we established in analyses I and II that *True Recurrence* does not affect the *Child Frequency / Adult Frequency* of low-frequency elements as much as with high-frequency elements, it is possible that the more low-frequency elements there are, the more elements will fall within a flat line at the bottom of the plot, while the high-frequency elements will form a slanted, upwards-going line. That may be why we see the strongest flat line with content words, a less obvious but still apparent flat line with function words, and no such line with POS bigrams.

2. Why do some elements have a higher chance of recurring than others?

Above and beyond a general frequency effect on the way in which the data pattern, we may pose the more fine-grained question of why it is that some elements have a higher *True Recurrence* than others. The plots suggest that this could be a frequency-effect as well: possibly, the more frequent an element, the higher its *True Recurrence*. The elements situated farther along the upwards-reaching part of the distributions, with a relatively large *True Recurrence*,

are certainly more frequent (= darker) than elements with a smaller *True Recurrence*. If this is a general pattern – one that is not just restricted to high-frequency data points –, then elements farther along the flat lines of low-frequency elements in Figures 17 (a) and (b) should *also* be more frequent than low-frequency elements that have a lower *True Recurrence* value. Perhaps, that is, frequency determines the *True Recurrence* of most linguistic elements, even though elements within the flat lines simply are not frequent enough for their high *True Recurrence* to have any effect on *Child Frequency / Adult Frequency*, which is why we find them within the flat lines at the bottom of the plots.

A full-scale investigation of the factors that cause some elements to be more likely to recur than others exceeds the scope of this thesis. But we can offer some suggestive evidence that there may be factors other than frequency which bring about an element's *True Recurrence* score. To procure such evidence, for each corpus and each element type, we select values for *True Recurrence* and *Child Frequency* that pick out two different groups of data points: (1) elements that are situated within the left half of the flat line at the bottom of the scatterplots and (2) elements that are situated within the right half of the line. For example, based on Figure 17 (a), we choose values for the two variables such that group (1) contains all elements with a *Child Frequency* lower than 0.3 and a *True Recurrence* lower than or equal to 0.25; for group (2), we choose values such that it contains all elements with a *Child Frequency* lower than 0.3 and a *True Recurrence* higher than 0.25. We can then inspect the average *Combined Child and Adult Frequency* of elements in the two groups. If it is true that elements with a higher frequency are more likely to recur, the frequency of elements in group (2) should be higher than that of elements in group (1).

Note that we choose the thresholds of 0.25 for *True Recurrence* and 0.3 for *Child Frequency* according to our intuitive understanding of the scatterplots and the patterns therein; and because the data take different values in different corpora, we choose different thresholds depending on which corpus we focus on. It may be possible to devise more sophisticated methods; but as our aim is not to perform an in-depth analysis, but rather to acquire a basic understanding of the relationship between frequency and *True Recurrence*, we contend ourselves with picking intuitively appropriate thresholds for all three corpora (again, cf. the appendix for plots based on the other two corpora). To keep things simple, we restrict analysis to scatterplots based on *True Recurrence* and *Child Frequency*; and since we find the largest number of low-frequency elements for content words, which increases the chance of producing significant results, we further restrict analysis to content words.

Table 17 shows the *True Recurrence* and *Child Frequency* thresholds we choose for the three corpora in order to split the data points on the low-frequency line at the bottom of the scatterplots into a left and right half; in addition, the table shows the mean *Combined Child and Adult Frequency* of the two groups, plus the results of a one-sided t-test comparing the two means. Surprisingly, the mean combined frequency of elements in the right half is lower than that of elements in the left half; and as the t-scores for Adam and Abe are statistically significant, this difference is unlikely to be due to chance. These results suggest that there must be something other than frequency that causes the relatively high *True Recurrence* of elements in the right half. After all, if frequency was the only determinant of *True Recurrence*, we should find the *opposite* pattern, with elements in the right half being *more* frequent.

Some indication as to what could be going on is given by the proportions of nouns, adjectives, and verbs in the two halves. As we make the transition from left to right half, the proportion of adjectives and verbs drops, and the proportion of nouns increases. In the Adam corpus, 61 % of the content words in the left half are nouns; in the Abe and Sarah corpora, this proportion is at 51 % and 60 %, respectively. In the right half, the proportions increase to 100 %, 76 %, and 90 %, respectively.

Potentially, a number of explanations could be offered for this pattern. Perhaps, nouns

Corpus	*True Rec.*	*Child Freq.*	Mean Freq. Left Half	Mean Freq. Right Half	t-score
Abe	0.25	0.3	0.34 (sd: 0.14, n = 58)	0.32 (sd: 0.10, n = 42)	0.98
Adam	0.4	1.0	1.13 (sd: 0.37, n = 104)	0.88 (sd: 0.29, n = 14)	2.42 *
Sarah	0.25	0.35	0.67 (sd: 0.32, n = 89)	0.47 (sd: 0.17, n = 64)	4.47 * * *

Table 17: Left and right halves of the flat lines at the bottom of the scatterplots based on content words. Figures in the second and third columns are the *True Recurrence* and *Child Frequency* thresholds for splitting the lines into a left and a right half. Figures in the fourth and fifth *mean frequency* columns are the mean *Combined Child and Adult Frequency* values for elements in the two halves. The t-scores in the sixth column pertain to the differences between those means.

somehow attract the interlocutors' attention, and so if they occur, they tend to be re-used more than other content words. And because most nouns are not very frequent, most nouns with a high probability of recurring are also infrequent. Or perhaps some low-frequency elements are generally unusual and therefore especially salient, and this is why interlocutors tend to re-use them – and because most low-frequency content words happen to be nouns, most content words with a high probability of recurring are also nouns. Without a more in-depth analysis, we can only speculate. It is clear, however, that while frequency appears to play an important role in bringing about the patterning of the data with respect to *True Recurrence*, there are hidden subtleties which hint at the involvement of factors other than just frequency.

6.3 Future Research and Open Questions

Apart from what causes some elements to have a higher chance of recurring than others, another issue that warrants further discussion is the design of a recurrence-quantificational study that properly utilizes the longitudinal nature of the CHILDES corpus. Related to that is the need for an evaluation of the general methodology – i.e., is it sensible to combine *Recurrence Quantificational Analysis* with the study of dialogue? Both points are addressed below.

1. Thoughts on a more longitudinally-oriented design

In the present thesis, we took a coarse-grained approach by taking averages over each corpus. That is, when speaking of the *True Recurrence* or *Child Frequency* of a single linguistic element, we were referring to its average *True Recurrence* or *Child Frequency*, calculated over all dialogue transcripts within the corpus in question. This approach lends itself to identifying robust, general patterns, and it is particularly suitable in light of the exploratory nature of this work. But with the general patterns now established, the next step should be the utilization of the longitudinal data in the CHILDES corpus to conduct a study that is actually longitudinal in nature. The results obtained via the general approach taken in this thesis are, in fact, well-suited for informing the design of a more longitudinal study. It may, in particular, be possible to build on the finding that *convergence* appears to have a much stronger impact on the usage of high-frequency elements than on the usage of low-frequency elements.

When new linguistic elements are introduced into the child's speech, it is conceivable that at least some of them at first start out as low-frequency elements, then to be continually reinforced through usage. In other words, some elements may start out as relatively infrequent elements; but over the course of acquisition, they may increase in frequency. If such a process can be observed, we may form the expectation that initially, while the element is not being used very frequently, *convergence* does not affect its usage. Then, as soon

as the element crosses a particular threshold of frequency, it may transition to the group of elements whose frequency is high enough for *convergence* to exert an effect on their usage, further strengthening its prevalence within future child-caregiver dialogues.

This hypothesized state of affairs could form the starting point for future research. If we take the measures used in this thesis to track *convergence* and frequency, and we calculate them on a per-dialogue basis, we can track how the frequency of elements changes over the course of development. If – as the results of this thesis suggest – *convergence* only exerts its effect on the frequency of a given element in case that element has a relatively high base-frequency to begin with, this should be reflected in how the frequency of elements with a high *True Recurrence* develops from dialogue to dialogue transcript.

Namely, if a low-frequency, high-recurrence element becomes more frequent, its frequency should at first increase relatively slowly, as *True Recurrence* does not exercise its full influence yet. But presumably, as soon as a given frequency threshold is crossed, we should begin to see its impact more, with possibly a sudden increase in frequency. This could be contrasted with low-frequency elements that are less likely to recur. Presumably, the frequency of such elements would increase more gradually, as *True Recurrence* would not reinforce their usage as much as with high-recurrence elements, even once a relatively high level of frequency has been reached.

More generally, we may imagine a state of affairs where, as the child matures, linguistic elements are introduced into the evolving child-caregiver dialogue, where they are either reinforced (selected for) and become integrated into future dialogues, or selected against, disappearing from or becoming less frequent in future dialogues. In reinforcing the usage of some but not other elements, *convergence* may play a crucial role in shaping not only future child-caregiver dialogues but also the more input-dependent parts of the language system the child is eventually going to develop. A longitudinal study based on some of the tools developed in the current thesis could help to shed light on the extent to which this is true or false.

2. Is Recurrence-Quantificational Analysis (RQA) a sensible approach?

RQA is a relatively complex methodology, and given the results obtained, it behooves us to reflect critically on its merits. Is *RQA* a necessary component of the analyses performed, or could we have used a simpler approach?

To address this question, we have to recognize that *RQA* is a set of tools for the comparison of two data series that is supposed to make this comparison palpable, or *simple* – or at least as simple as possible. In order to properly compare the two sequences in their entirety, we need to compare every data point in the first to every data point in the second sequence; otherwise, we will neglect some part of the possible total comparison of the two series. But this is, essentially, what a recurrence plot does: it compares all possible pairs of data points and arranges them, in a certain order, on a two-dimensional coordinate system. We can choose to visualize this coordinate system in some way or another, but we do not have to. Thus, recurrence plots embody an essential requirement for comparing two series of data points – and no less so if the two series consist of conversational turns. It is, therefore, definitely sensible to consider *RQA* when we are interested in comparing sequences of turns in dialogue.

A possible alternative would be to work with raw orderings of pairs of turns, which arguably would be more basic data structures than recurrence plots. But recurrence plots offer additional advantages. By arranging the pairs in a certain way, we are able to identify certain areas within the plots with specific types of pairs – for example, all the pairs of turns where the turn by speaker A precedes or follows the turn by speaker B, or all the pairs of turns that are separated by at most n other turns. Since these types of pairs are located

within specific sub-regions of a recurrence plot, they become synonymous with geometrical distinctions. If we worked with raw orderings of turns, we would have to identify such pairs of turns via some other, non-geometrical mechanism. Perhaps it is possible to do so in an intuitively appealing way. But the fact remains that *RQA* already offers such a way. It offers, quite possibly, the best available tools for comparing sequences of turns, and we see no reason to think up a different approach – which would likely be equivalent on an abstract level anyway.

7 Conclusion

This thesis investigated the impact of two separate influences on the frequency with which words and syntactic structures are used by child and adult: (1) their frequency in the other interlocutor's speech and (2) the frequency with which they are used by child and adult in temporally close turns. With some support from previous work (Dale & Spivey, 2006; Fernández & Grimm, 2014), the latter was assumed to track local adaptation mechanisms whereby child and adult match one another's language use. Such local mechanisms have been suggested to be partly responsible for the shaping of child-directed speech (CDS) – a unique, simplified type of speech adults often use when speaking with young children (Snow, 1995; Kunert et al., 2011). But of course, local adaptation – or *convergence*, as we have called it – is unlikely to only influence the adult's speech (Veneziano & Parisse, 2010). Consequently, the focus of this thesis was on how both the adult's *and* the child's speech are affected.

The results are strongly indicative of *convergence* playing an important role in shaping the speech of both interlocutors. When considering *all* linguistic elements, frequency in the other's speech is the stronger determinant for the majority of element types; but crucially, as soon as we consider only high-frequency words and structures, the impact of *convergence* increases so much that out of the two determinants, it can be considered the dominant overall factor in shaping the language use of the other speaker. Because the most frequent constituents account for the largest part of the actual language produced, the results support a possible conversational dynamics where child and caregiver adjust their speech to one another largely through local adaptation mechanisms. This has implications for the formation of CDS in that it stresses the importance of *convergence*, but it also raises the possibility that *convergence* may be involved in the child's acquisition of new words and syntactic structures.

Two other notable findings have to do with the homogeneity of patterns across corpora and interlocutors, as well as with the importance of child-adaptation to the adult. As far as the former is concerned, we find remarkable cross-corpora consistency of the influence of *convergence* and frequency on child and adult speech, but also some individual variation, particularly when comparing the Sarah corpus to the other two corpora. We also find that the way in which the adult's speech is affected is very close to the way in which the child's speech is affected by the two factors. Overall, the results suggest that despite some individual variation, caregiver-child dyads may match their language use in similar ways; and possibly, the adult may match the child's speech in a manner similar to how the child matches the adult's speech. A likely difference in how the interlocutors adapt to one another's language use is highlighted by the relative importance of child-adaptation to the adult, which compared to adult-adaptation to the child clearly is the stronger predictor of usage. That is, the way in which the child adapts to the adult shapes both interlocutors' language use more strongly than the way in which the adult adapts to the child. This suggest that the child may model her speech more directly on preceding adult turns, whereas the caregiver may actually adapt more strongly to the way in which the child adapts to *his* utterances.

But apart from these more specific dynamics, the overall pattern discovered in the present thesis is one where usage is extensively affected by local adaptation. On a conservative interpretation, one would probably have to consider the effect of *convergence* to be on equal footing with the effect of general frequency. And if one chooses to focus only on words and syntactic structures with a relatively high frequency, one would have to conclude that the impact of *convergence* is, in fact, the stronger determinant of usage. This thesis has thus shown that CDS is, at least in part, the result of local adaptation mechanisms, but also that part of the child's language is likely the result of the very same adaptation process.

References

Aksu-Koç, A. (2006). *The acquisition of aspect and modality: The case of past reference in turkish*. Cambridge: Cambridge University Press.

Angus, D., Smith, A., & Wiles, J. (2012). Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics*, *18*(6), 988–997.

Brennan, & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Brown, R. (1973). *A first language: The early stages.* Cambridge, Massachusetts: Harvard University Press.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in korean: A cross-linguistic study. *Journal of child language*, *22*, 497–530.

Dale, R., & Spivey, M. (2005). Categorical recurrence analysis of child language. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual meeting of the cognitive science society* (pp. 530–535). Mahwah, New Jersey: Erlbaum.

Dale, R., & Spivey, M. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*(3), 391–430.

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In F. Keller & D. Reitter (Eds.), *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics* (pp. 76–87). Portland, Oregon: Association for Computational Linguistics.

DiDonato, M. D., England, D., Martin, C. L., & Amazeen, P. G. (2013). Dynamical analyses for developmental science: A primer for intrigued scientists. *Human Development*, *56*(1), 59–75.

Eckmann, J.-P., Kamphorst, S. O., & Ruelle, D. (1987). Recurrence plots of dynamical systems. *EPL (Europhysics Letters)*, *4*(9), 973.

Fernández, R., & Grimm, R. (2014). *Quantifying categorical and conceptual convergence in child-adult dialogue.* (To appear in Proceedings of the 36th annual conference of the cognitive science society. Quebec City, Canada)

Gao, J., & Cai, H. (2000). On the structures and quantification of recurrence plots. *Physics Letters A*, *270*(1), 75–87.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

Giles, H., Coupland, N., & Coupland, I. (1991). 1. accommodation theory: communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation.* Cambridge: Cambridge University Press.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the ieee conference on acoustics, speech, and signal processing* (Vol. 1, pp. 517–520). San Francisco, California.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, *37*(1), 3–19.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*(1), 39–44.

Kuczaj, I., & Stan, A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 589–600.

Kunert, R., Fernández, R., & Zuidema, W. (2011). Adaptation in child directed speech: Evidence from corpora. In R. Artstein, M. Core, D. DeVault, K. Georgila, E. Kaiser, & A. Stent (Eds.), *Proceedings of the 15th semdial workshop on the semantics and pragmatics of dialogue* (pp. 112–119). Los Angeles, California: Institute for Creative Technologies, University of Southern California.

Leonardi, G. (2012). The study of language and conversation with recurrence analysis methods. *Psychology of Language and Communication*, *16*(2), 165–183.

Levelt, W. J., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*(1), 78–106.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, *26*(4), 657–657.

Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*(5), 299–307.

Newport, E., Gleitman, H., & Gleitman, L. R. (1977). Mother, i'd rather do it myself: Some effects and noneffects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children* (pp. 109–49). Cambridge: Cambridge University Press.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169–190.

Pinker, S. (1994). *The language instinct: The new science of language and mind* (Vol. 7529). Londong: Penguin UK.

Reitter, D., Moore, J. D., & Keller, F. (2010). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In R. Sun (Ed.), *Proceedings of the 28th annual conference of the cognitive science society.* Mahwah, New Jersey: Erlbaum.

Saxton, M. (2009). The inevitability of child directed speech. In S. Foster-Cohen (Ed.), *Advances in language acquisition* (pp. 62–86). London: Palgrave Macmillan.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319.

Snow, C. (1989). Understanding social interaction and language acquisition; sentences are not enough. In M. H. Bornstein & J. S. Bruner (Eds.), *Interaction in human development* (pp. 62–86). Hillsdale, New Jersey: Erlbaum.

Snow, C. (1995). Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 180–193). Hoboken, New Jersey: Blackwell.

Sokolov, J. L. (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, *29*(6), 1008.

Tardif, T., Shatz, M., & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of english, italian, and mandarin. *Journal of Child Language*, *24*(3), 535–565.

Veneziano, E., & Parisse, C. (2010). The acquisition of early verbs in french: Assessing the role of conversation and of child-directed speech. *First Language*, *30*(3-4), 287–311.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of linguistics*, *19*(1), 29–58.

A Appendix

Listed below are, for every corpus, scatterplots showing *Child Frequency* and *Adult Frequency*, respectively, as a function of *True Recurrence*. Data points are colored according to their *Combined Child and Adult Frequency*, with a darker color representing a higher value. The data are presented in batches of nine scatterplots each. From left to right, the plots in each column of a given batch show the data for content words, function words, and POS bigrams. From top to bottom, each row contains plots showing all data points, the data from the high-frequency group, and the data from the low-frequency group. The scatterplots in the second (high-frequency group) and third (low-frequency group) row additionally contain linear regression lines fitted to the data within individual plots.

A.1 True Recurrence and Child Frequency

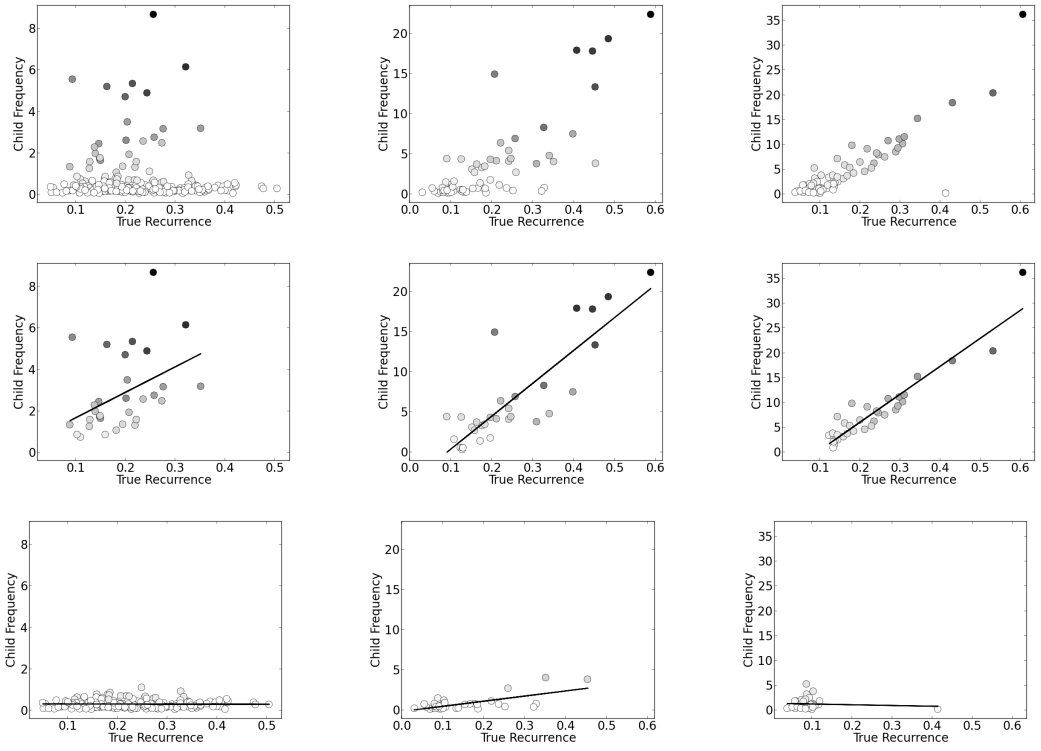

Figure 18: Abe corpus. *Child Frequency* as a function of *True Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

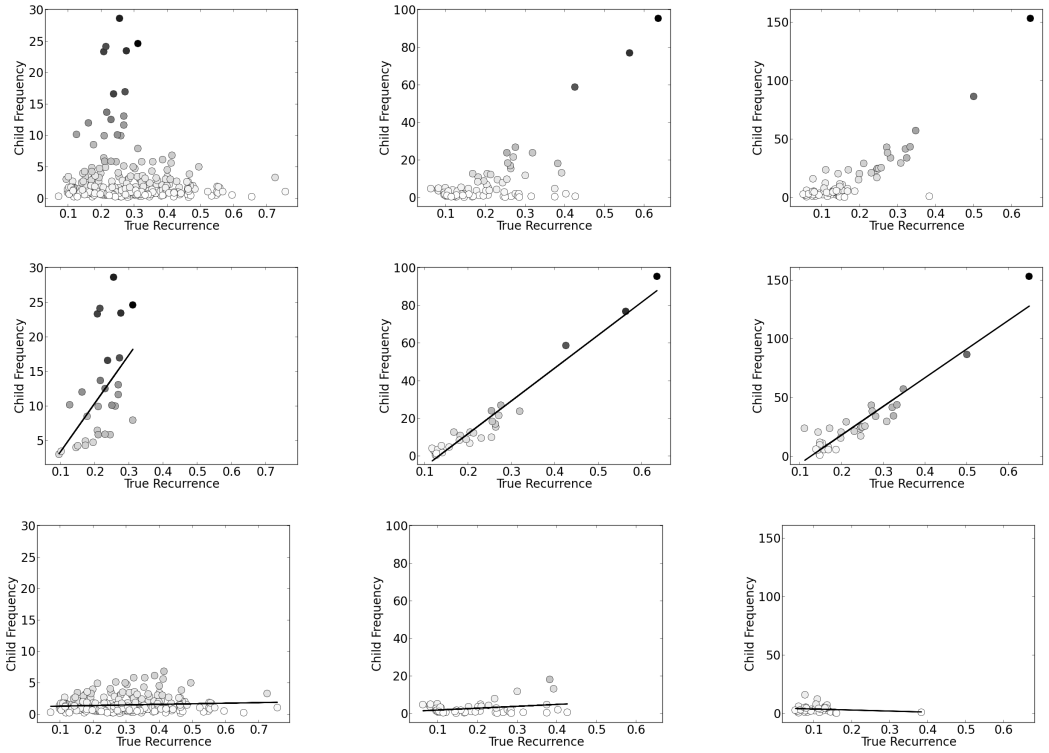

Figure 19: Adam corpus. *Child Frequency* as a function of *True Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

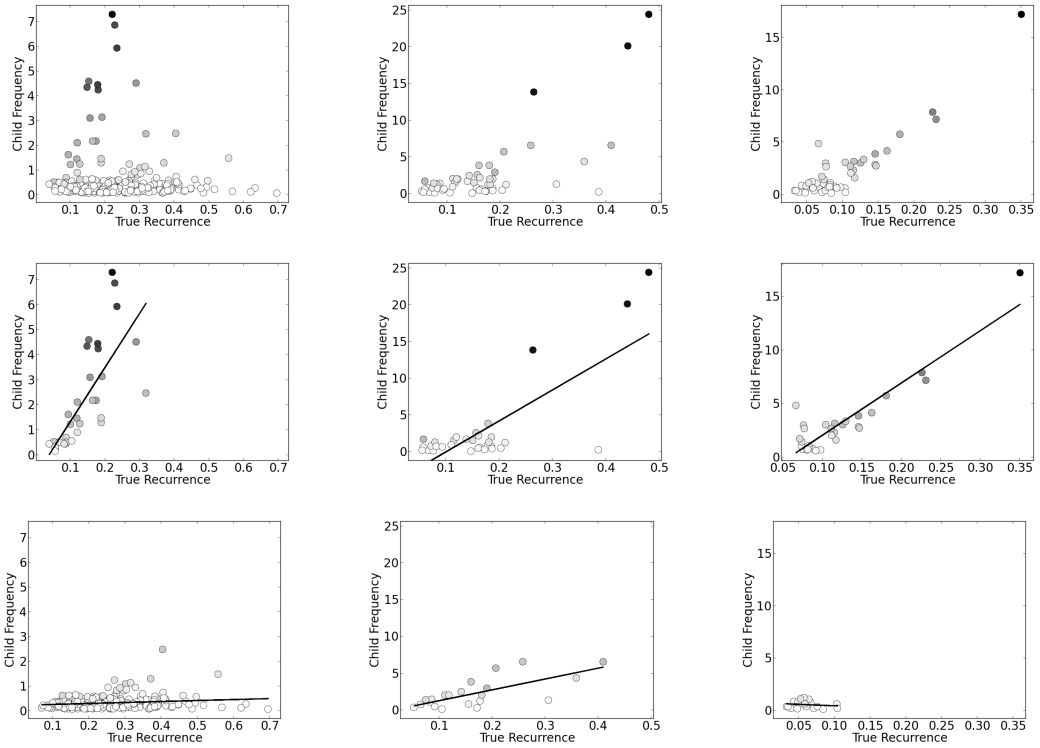

Figure 20: Sarah corpus. *Child Frequency* as a function of *True Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

A.2 True Recurrence and Adult Frequency

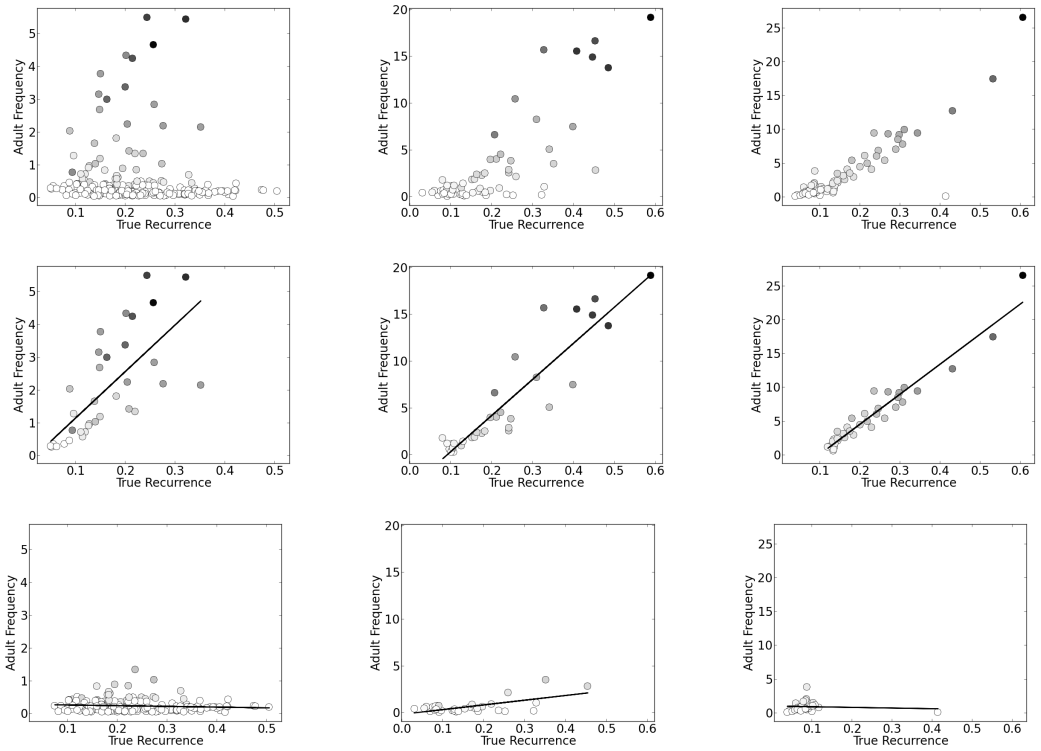

Figure 21: Abe corpus. *Adult Frequency* as a function of *True Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

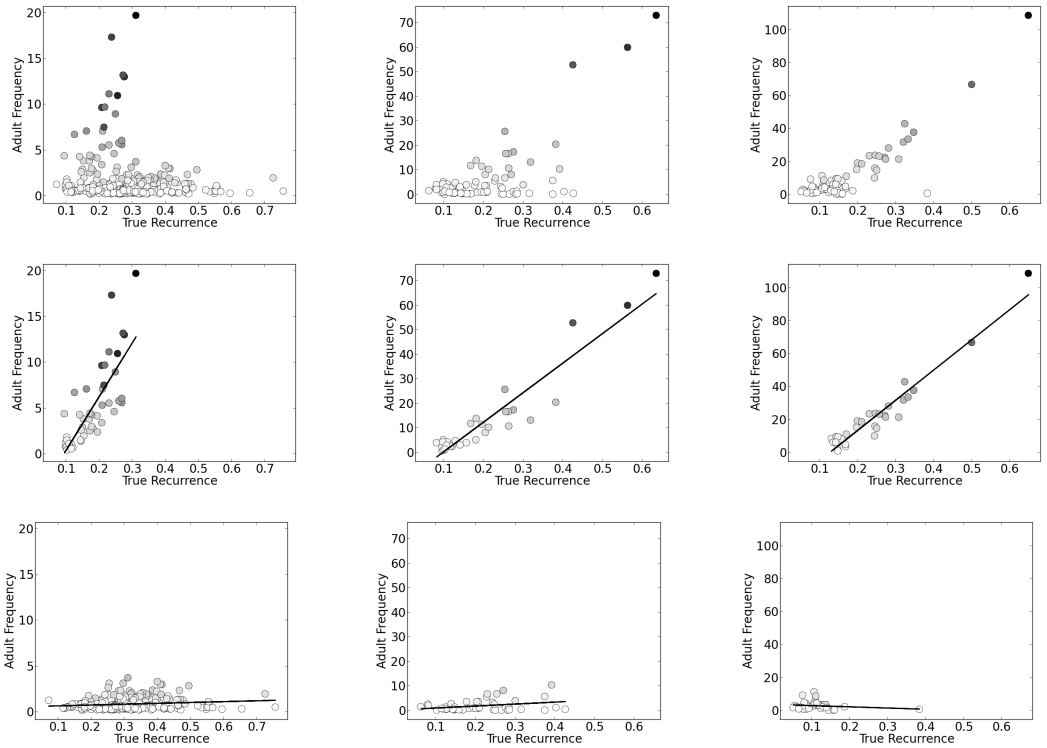

Figure 22: Adam corpus. *Adult Frequency* as a function of *True Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

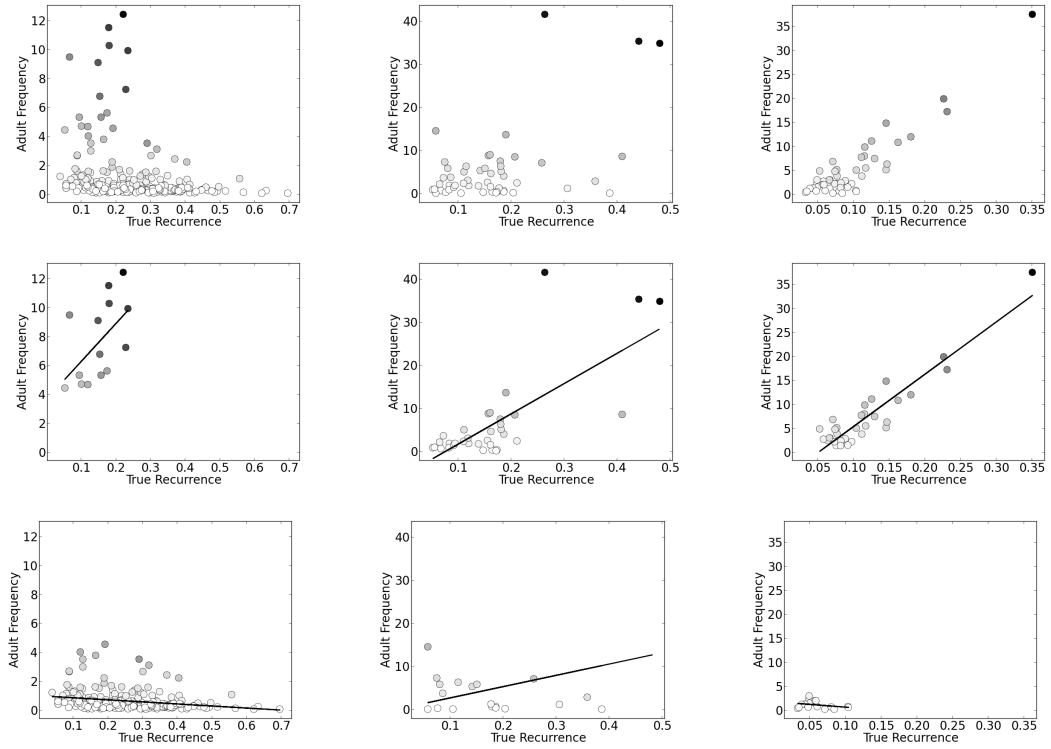

Figure 23: Sarah corpus. *Adult Frequency* as a function of *True Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

A.3 True Adult-Initiated Recurrence and Child Frequency

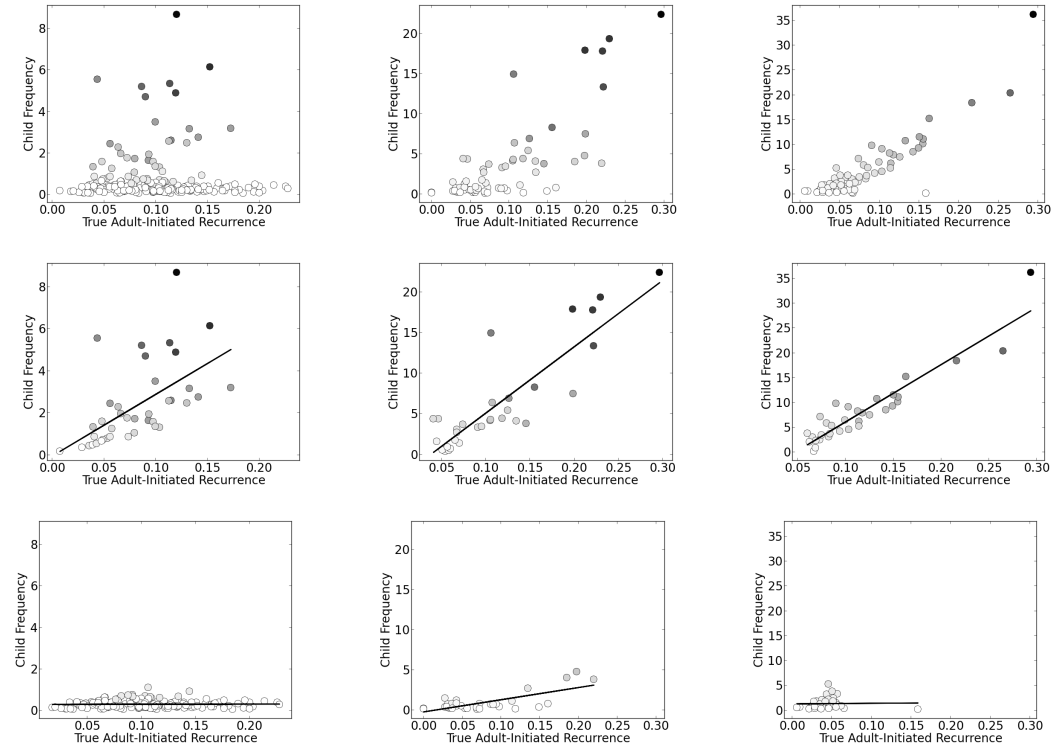

Figure 24: Abe corpus. *Child Frequency* as a function of *True Adult-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

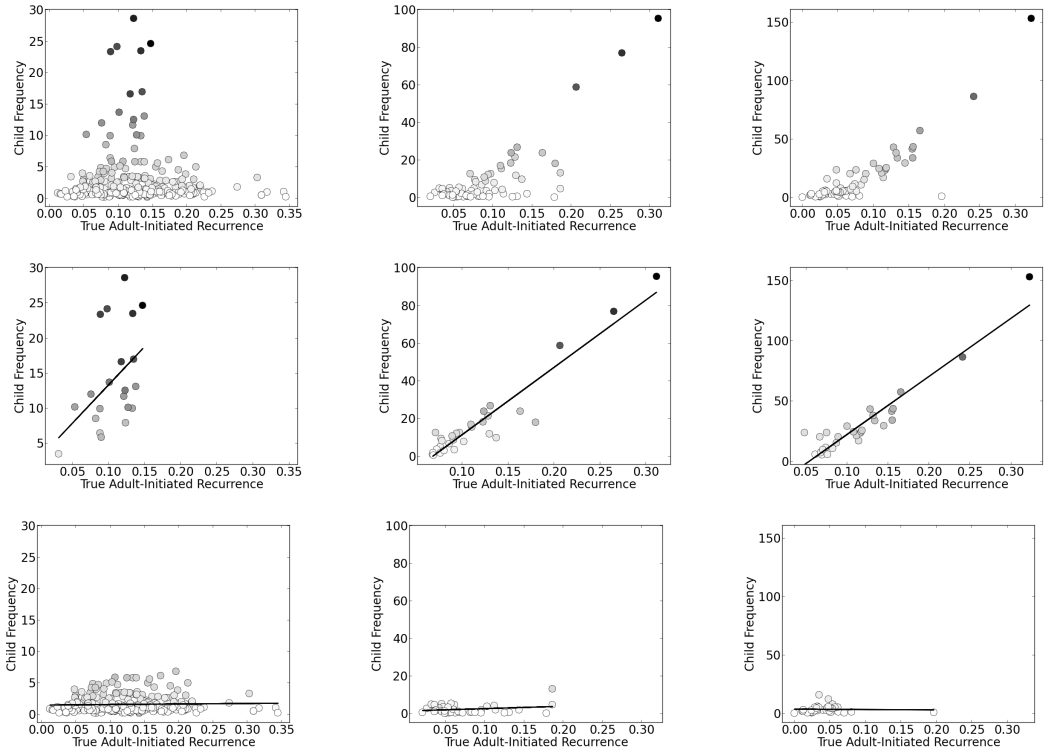

Figure 25: Adam corpus. *Child Frequency* as a function of *True Adult-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

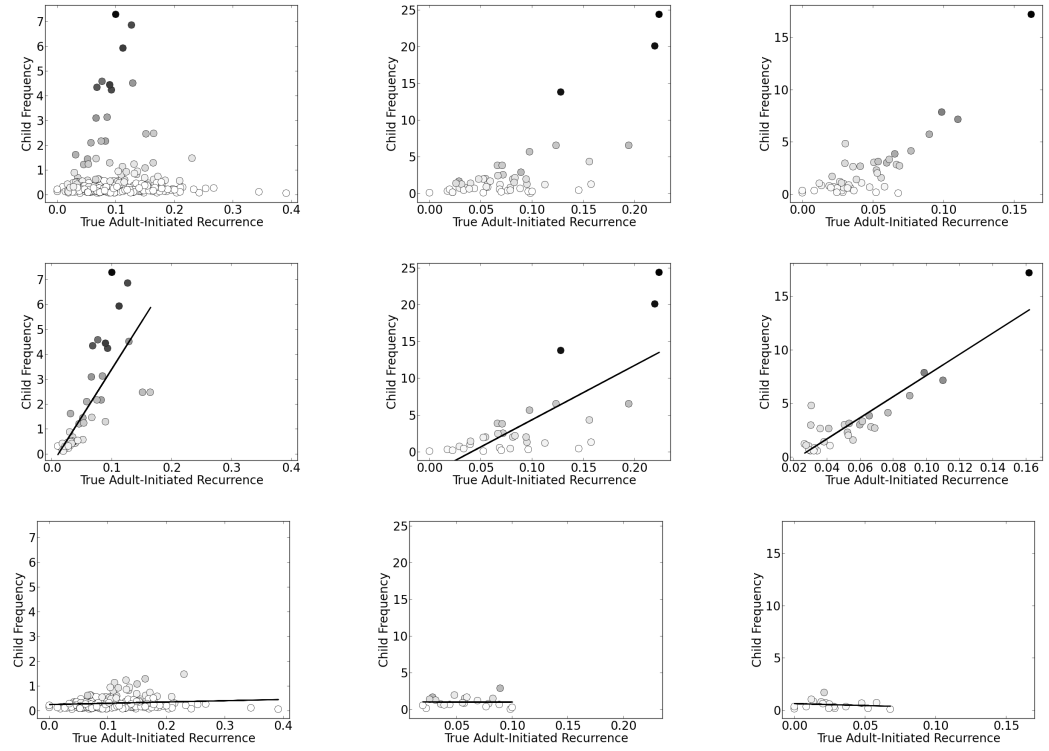

Figure 26: Sarah corpus. *Child Frequency* as a function of *True Adult-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

A.4 True Adult-Initiated Recurrence and Adult Frequency

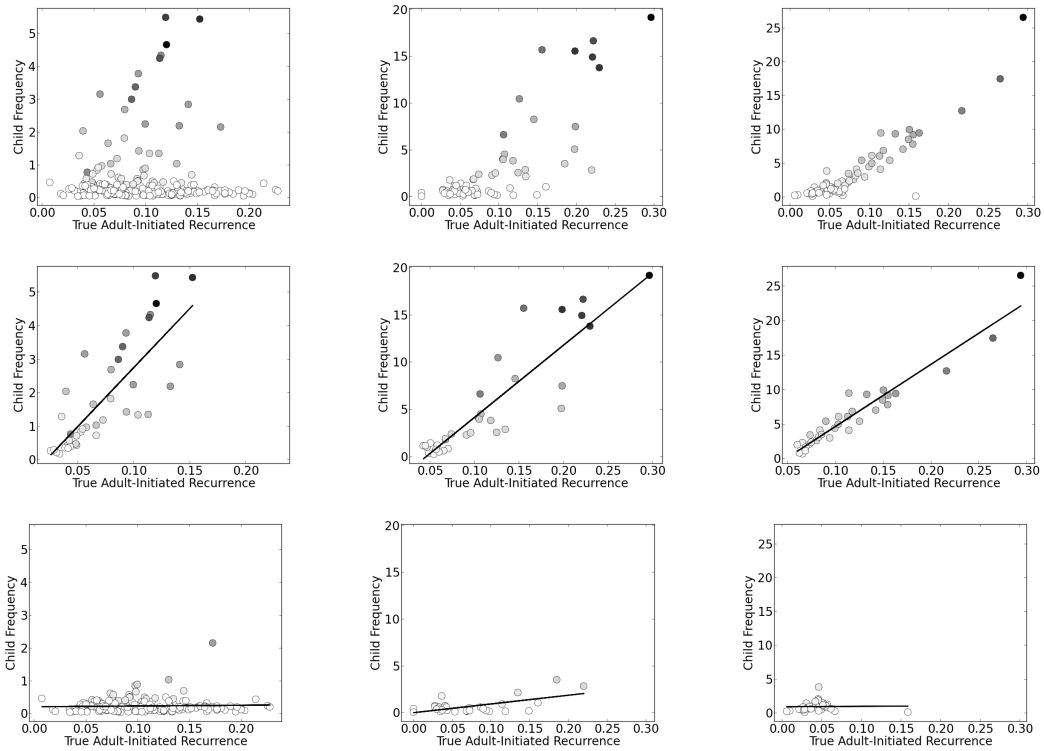

Figure 27: Abe corpus. *Adult Frequency* as a function of *True Adult-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

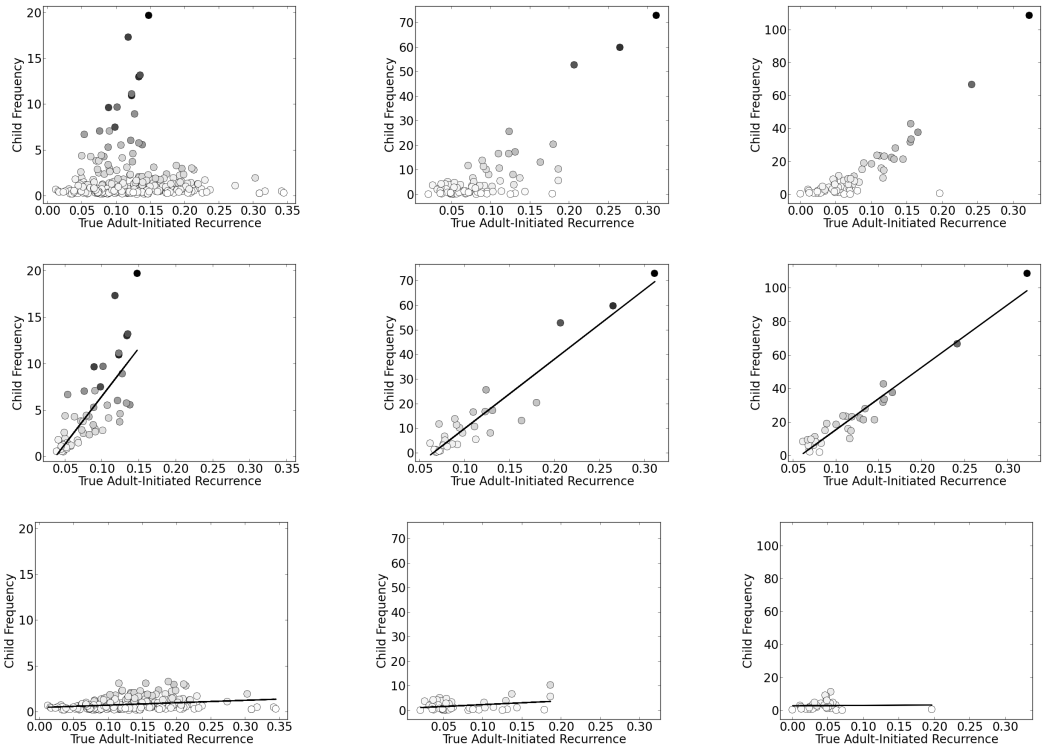

Figure 28: Adam corpus. *Adult Frequency* as a function of *True Adult-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

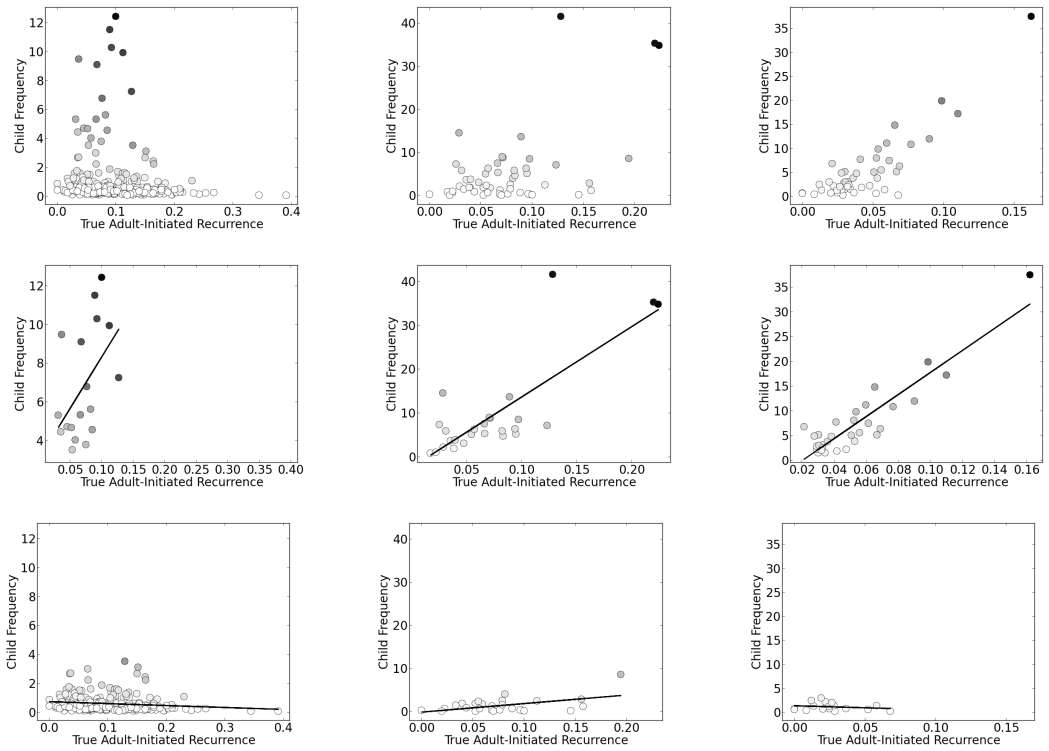

Figure 29: Sarah corpus. *Adult Frequency* as a function of *True Adult-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

A.5 True Child-Initiated Recurrence and Child Frequency

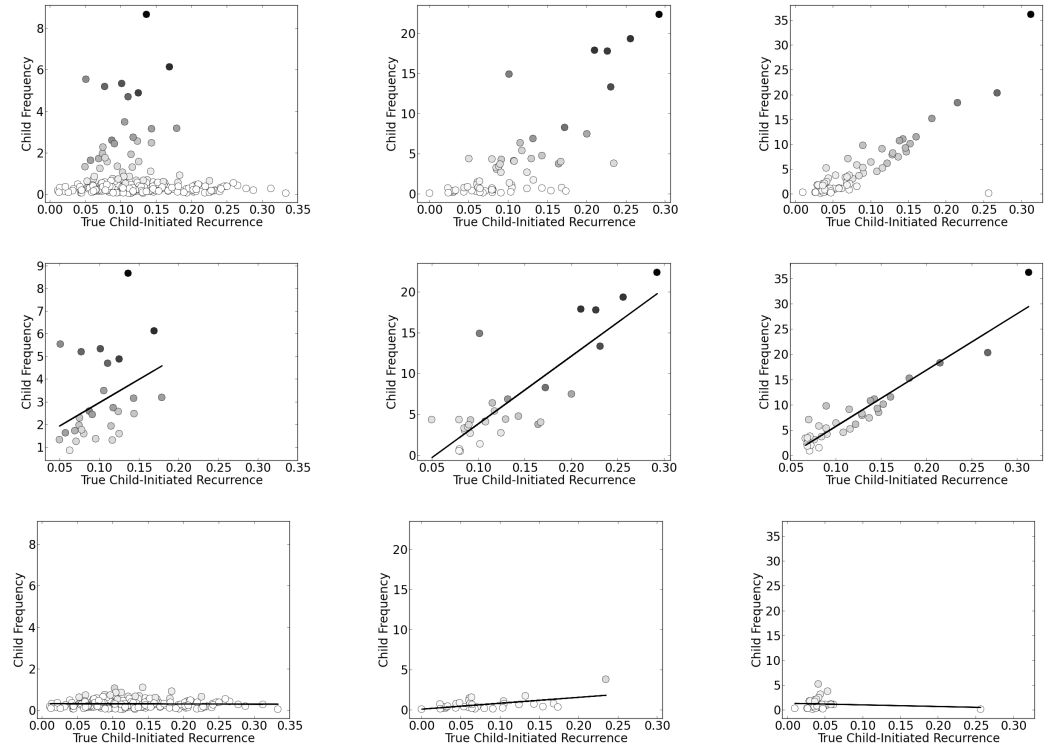

Figure 30: Abe corpus. *Child Frequency* as a function of *True Child-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

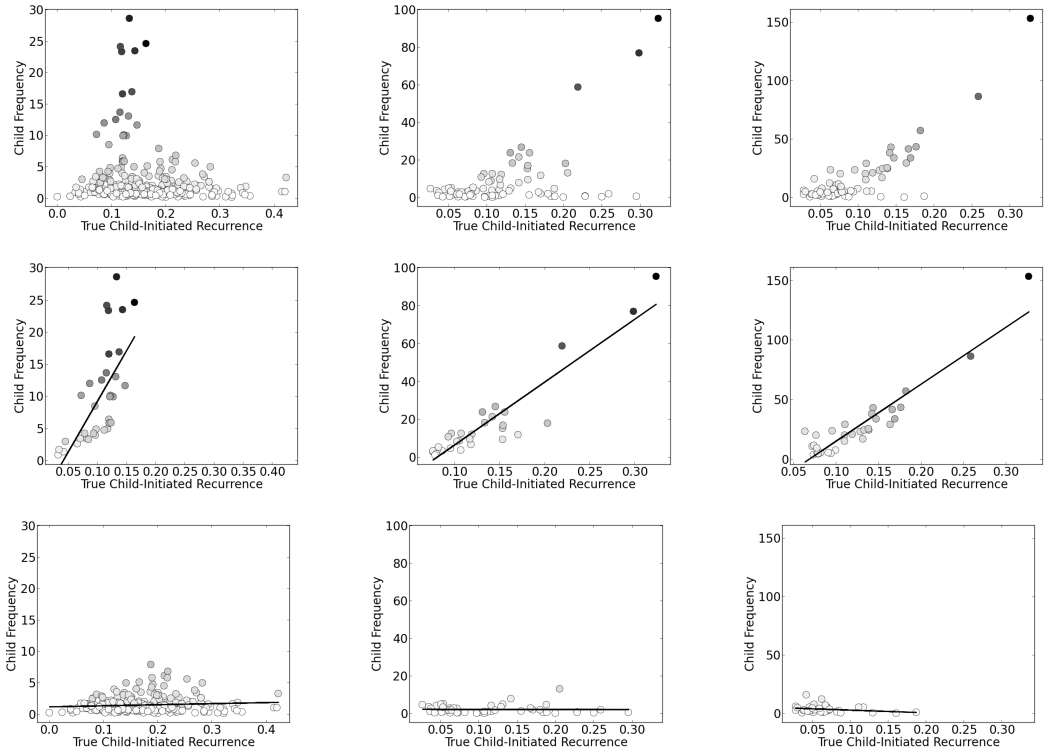

Figure 31: Adam corpus. *Child Frequency* as a function of *True Child-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

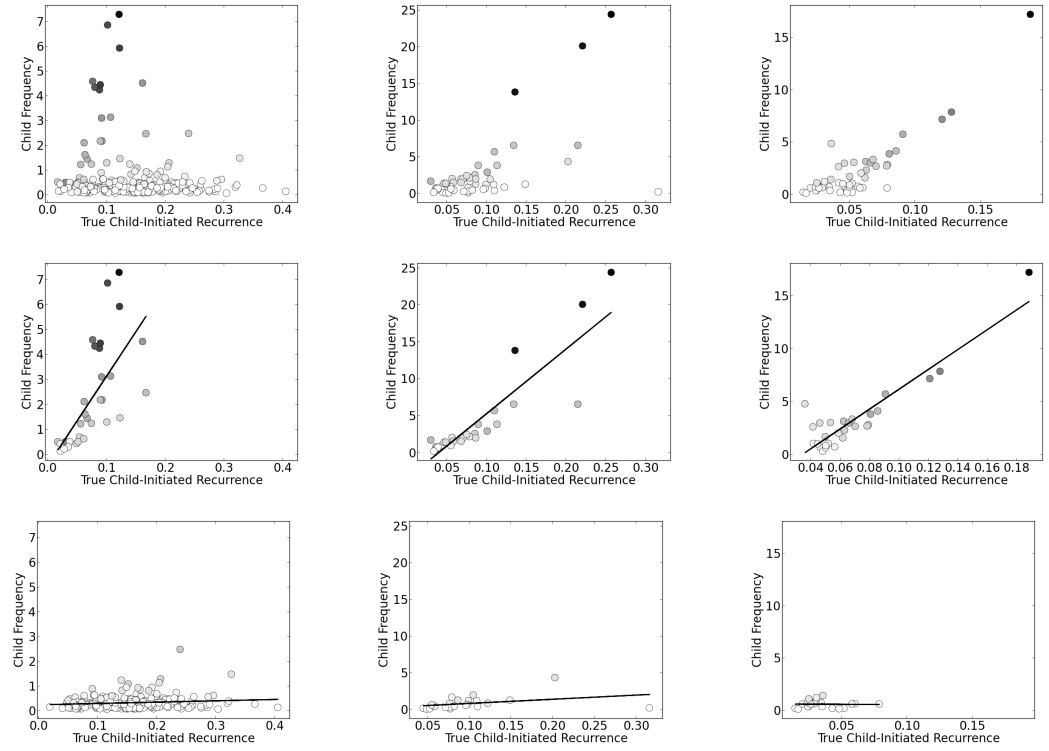

Figure 32: Sarah corpus. *Child Frequency* as a function of *True Child-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

A.6 True Child-Initiated Recurrence and Adult Frequency

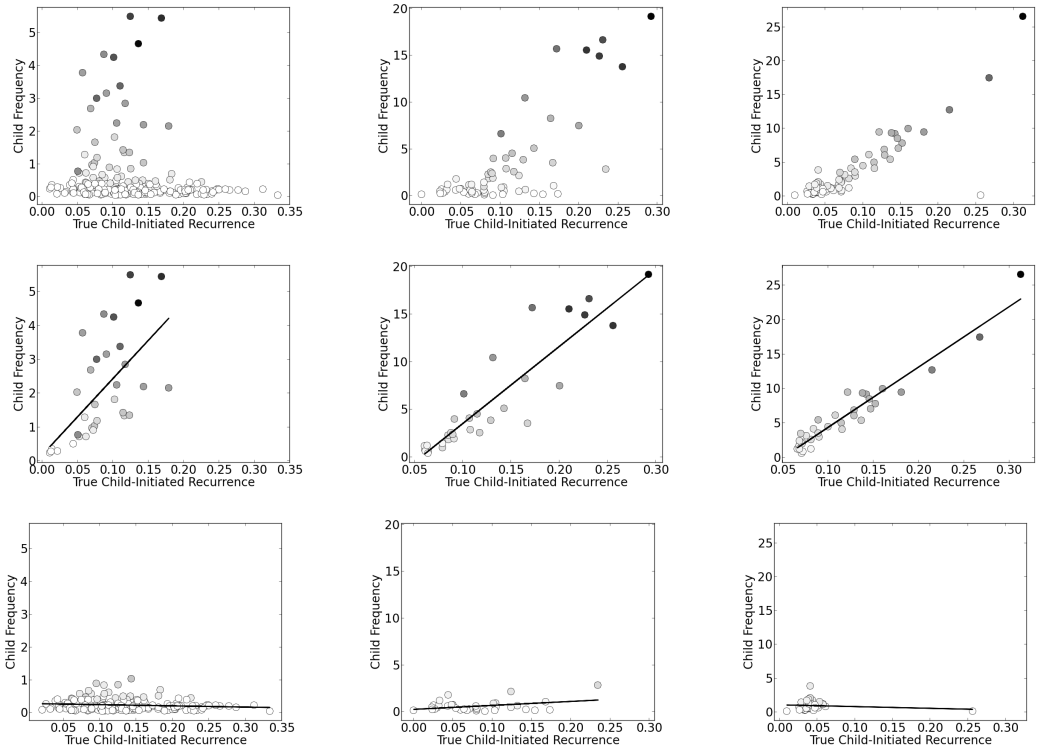

Figure 33: Abe corpus. *Adult Frequency* as a function of *True Child-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

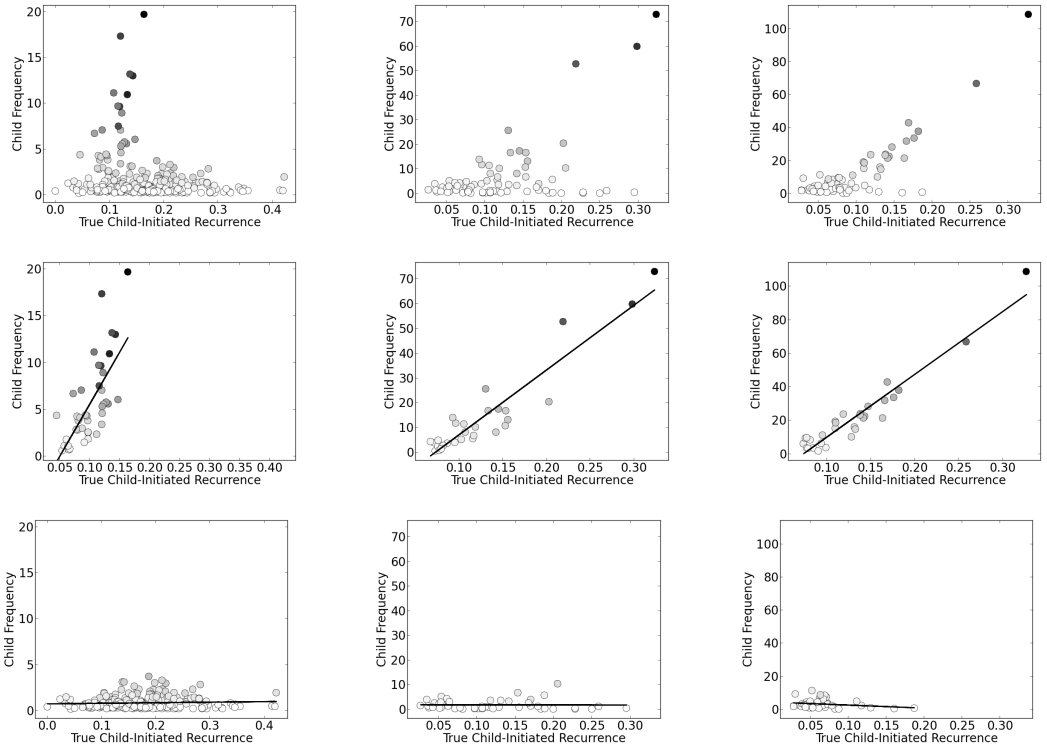

Figure 34: Adam corpus. *Adult Frequency* as a function of *True Child-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group

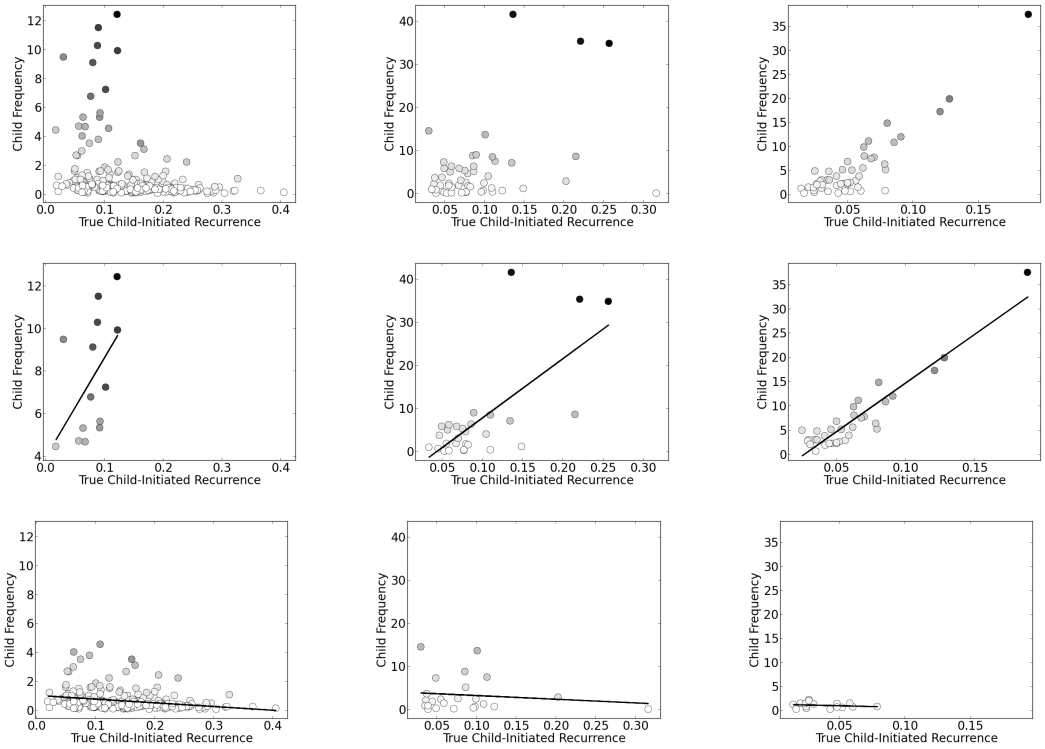

Figure 35: Sarah corpus. *Adult Frequency* as a function of *True Child-Initiated Recurrence*.
Left: Content Words – Middle: Function Words – Right: POS Bigrams
Top: All Data Points – Middle: High-Frequency Group – Bottom: Low-Freq. Group