Cognition 221 (2022) 104986



Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit





Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages

Nicholas A. Lester [a,*], Steven Moran [a,b,*], Aylin C. Küntay [c], Shanley E.M. Allen [d], Barbara Pfeiler [e], Sabine Stoll [a]

[a] Department of Comparative Language Science & Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Thurgauer Strasse 30, 8050 Zürich, Switzerland
[b] Institute of Biology, University of Neuchâtel, Rue Emile-Argand 11, CH-2000, Neuchâtel, Switzerland
[c] Department of Psychology, Koç University, Rumelifeneri Yolu, Sarıyer, 34450 İstanbul, Turkey
[d] Psycholinguistics and Language Development Group, Department of Social Sciences, University of Kaiserlautern, TU Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany
[e] National Autonomous University of Mexico, Centro Peninsular en Humanidades y Ciencias Sociales, Ex Sanatorio Rendón Peniche, Calle 43 s/n entre 44 y 46, col. Industrial, 97150 Mérida, Yucatán, Mexico

ARTICLE INFO

Keywords:
Cross-linguistic language acquisition
Variation sets
Input patterns
Child-directed speech

ABSTRACT

Caretakers tend to repeat themselves when speaking to children, either to clarify their message or to redirect wandering attention. This repetition also appears to support language learning. For example, words that are heard more frequently tend to be produced earlier by young children. However, pure repetition only goes so far; some variation between utterances is necessary to support acquisition of a fully productive grammar. When individual words or morphemes are repeated, but embedded in different lexical and syntactic contexts, the child has more information about how these forms may be used and combined. Corpus analysis has shown that these partial repetitions frequently occur in clusters, which have been coined variation sets. More recent research has introduced algorithms that can extract these variation sets automatically from corpora with the goal of measuring their relative prevalence across ages and languages. Longitudinal analyses have revealed that rates of variation sets tend to decrease as children get older. We extend this research in several ways. First, we consider a maximally diverse sample of languages, both genealogically and geographically, to test the generalizability of developmental trends. Second, we compare multiple levels of repetition, both words and morphemes, to account for typological differences in how information is encoded. Third, we consider several additional measures of development to account for deficiencies in age as a measure of linguistic aptitude. Fourth, we examine whether the levels of repetition found in child-surrounding speech is greater or less than what would have been expected by chance. This analysis produced a new measure, redundancy, which captures how repetitive speech is on average given how repeititive it could have been. Fifth, we compare rates of repetition in child-surrounding and adult-directed speech to test whether variation sets are especially prevalent in child-surrounding speech. We find that (1) some languages show increases in repetition over development, (2) true estimates of variation sets are generally lower than or equal to random baselines, (3) these patterns are largely convergent across developmental indices, and (4) adult-directed speech is reliably less redundant, though in some cases more repetitive, than child-surrounding speech. These results are discussed with respect to features of the corpora, typological properties of the languages, and differential rates of change in repetition and redundancy over children's development.

* Corresponding authors.
E-mail addresses: nicholas.a.lester@gmail.com (N.A. Lester), steven.moran@uzh.ch (S. Moran), akuntay@ku.edu.tr (A.C. Küntay), allen@sowi.uni-kl.de (S.E.M. Allen), sabine.stoll@uzh.ch (S. Stoll).

https://doi.org/10.1016/j.cognition.2021.104986
Received 11 February 2020; Received in revised form 16 August 2021; Accepted 5 December 2021
Available online 23 December 2021
0010-0277/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

One of the unresolved questions of language learning is how infants can extract and generalize linguistic units from the speech to which they are exposed. A diverse and growing body of research has begun to tease apart what aspects of the input could impact language development, with evidence coming from naturalistic (e.g., Aguado-Orea, 2004; Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002; Krajewski, Lieven, & Theakston, 2012), experimental (e.g., Branigan & Messenger, 2016; Lieven & Stoll, 2013; Savage, Lieven, Theakston, & Tomasello, 2006; Vasilyeva & Waterfall, 2012), computational (e.g., Freudenthal, Pine, Aguado-Orea, & Gobet, 2007; Vogt & Lieven, 2010), and mixed-methodological studies (e.g., Naigles & Hoff-Ginsburg, 1998). Nevertheless, there are still large gaps in our understanding of how input to children is structured, and in particular how the input may differ cross-linguistically.

Perhaps the most well-established link between input and acquisition is repetition. In general, languages are more repetitive than not (Haiman, 1997; Jakobson, 1966), and this repetitiveness has been shown to support language learning (Ambridge, Kidd, Rowland, & Theakston, 2015; Bannard & Lieven, 2009; Bard & Anderson, 1983; Brown, 1999; Cameron-Faulkner, Lieven, & Tomasello, 2003; Hoff-Ginsberg, 1986, 1990; Horst, Parsons, & Bryan, 2011). In fact, one of the best predictors of learning across levels of linguistic structure is pure frequency (e.g., Ambridge, Kidd, Rowland, & Theakston, 2015). For example, words that appear more often in the input are learned earlier (e.g., Braginsky, Yurovsky, Marchman, & Frank, 2016;Goodman, Philip, & Li, 2008). But frequency effects are also found beyond individual words. Repetition in multi-word contexts (contiguous or non-contiguous) offers reliable cues not only to word segmentation and meaning (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) but also to more general categories, such as word class (Cameron-Faulkner et al., 2003; Cartwright & Brent, 1997; Gómez & Maye, 2005; Mintz, 2003, 2006; Moran et al., 2018; Redington, Chater, & Finch, 1998; Santelmann & Jusczyk, 1998; Stoll, Abbot-Smith, & Lieven, 2009). Thus, repetition in the input provides multiple sources of information about the building blocks of language, and children appear quite able to incorporate this information during their early linguistic development.

Repetition has a natural counterpart: variability. Assuming a large enough textual window, each instance of a given word will occur in a different context, i.e., language is non-stationary (see Jurafsky & Martin, 2008: Ch. 12). Like repetition, variability has consequences for language processing and learning. For example, greater variability in the aggregate contextual distributions of words has been shown to facilitate comprehension (e.g., ; Moscoso del Prado Martín, Kostić, & Baayen, 2004) and acquisition (Lester, 2018; Waterfall, 2006), independent of pure frequency. Therefore, repetition and variability work hand in hand to support online processing in the short term and acquisition over the long term. The question is whether, and if so how speech to children is structured to support these pathways for learning.

In the present study, we focus on one facet of child-directed speech that embodies both repetition and variability, namely variation sets (Küntay & Slobin, 1996). Variation sets are groups of partially repetitive utterances that are tightly clustered in time. Crucially, one or more words are repeated but in different morpho-syntactic contexts. Küntay and Slobin (1996) provide the following example:

(1)
Who did we **see** when we went out shopping today?
Who did we **see**?
Who did we **see** in the store?
Who did we **see** today?
When we went out shopping, who did we **see**?

In this interaction a father prompts the memory of his child (aged 2;3) through a number of consecutive utterances focusing on the same topic of speech. The verb *see* is repeated five times and embedded in five different syntactic contexts.

Sequences like this one serve several purposes, each of which could support acquisition. For one, they orient and maintain the child's attention on a circumscribed topic (here, the memory of seeing someone at the store). They also promote comprehension by including new information (the variable content) piecemeal in relation to a lexical or conceptual point fo stability. Finally, as Küntay and Slobin (1996) argue, they could help children to track lexical roots across variable morphological and syntactic structures, thereby providing abundant cues to the formal and functional properties of the repeated element.

Waterfall (2006) explicitly tested whether the frequency of variation sets contributes to learning in a corpus of naturalistic adult-child interactions. In line with the literature cited above, she finds that words that occur more frequently in variation sets are produced earlier by young children. Computational research further supports this finding. Frank, Tenenbaum, and Fernald (2013) find that repetition of object labels in successive utterances leads to reasonably reliable classification of the intended referent (hence acquisition of the form-meaning mapping) above and beyond other more obvious cues, such as eye gaze and pointing. Experimental work further supports facilitation of learning due to partial repetitions across successive utterances. Schwab and Lew-Williams (2016, 2017) find that young children only successfully acquire novel object labels when these labels are repeated in blocks of successive sentences. Compatible effects were reported by Onnis, Waterfall, and Edelman (2008) for adults in an artificial language learning paradigm. Participants were better at detecting word and phrase boundaries in a novel language when learning trials were grouped with some degree of repetition across adjacent stimuli. Finally, as also pointed out by Onnis et al. (2008), the compressed temporal profile of variation sets allows even memory-limited learners to compare and discover structure in their input. Taken together, these findings suggest that where variation sets are present, either naturally or through experimental manipulation, learning outcomes improve across a number of linguistic levels.

But how prevalent are variation sets in speech? Presumably, for them to make a substantial impact on language acquisition, variation sets should appear in children's earliest interactions with a fair degree of regularity. Recent advances in the computational literature have demonstrated that prevalence of variation sets can be reliably estimated using automatic extraction techniques. Once the variation sets are extracted, prevalence is typically defined as the proportion of utterances (out of all utterances in the sample) that belong to at least one variation set. Applying these techniques, several studies have found evidence of variation sets in child-directed speech sampled from many different languages (Grigonytė & Björkenstam, 2016; Hoiting & Slobin, 2002; Küntay & Slobin, 1996; Küntay & Slobin, 2002; Waterfall, 2006; Wirén, Kristina, Björkenstam, & Cortes, 2016). These techniques have also been used to examine how the prevalence of variation sets changes as children mature. Results from these studies show that the vast majority of languages show a decrease in the prevalence of variation sets as children grow older (the others showing no change over time; Grigonytė & Björkenstam, 2016; Wirén et al., 2016). The evidence thus points to a strong link between the child's development and the caretakers' use of repetition. In the early stages, the caretaker speech is more repetitive, which supports learning at a time when children manifest fewer signs of comprehension and/or productive ability. Over time, as the children become more proficient language users, the caretaker speech becomes less repetitive, making room for more varied, contentful, and efficient communication.

It follows from this discussion that variation sets could be a universal feature of adult-child interactions, arising and subsiding in response to the evolving pressures of adult-child interaction, supporting acquisition all the while. However, the available evidence suffers from some shortcomings that must be addressed before we can draw such a bold conclusion.

First, the sample of languages that have been explored so far has not been adequately controlled for cultural, geographic, and linguistic

diversity. Specifically, prior research has either (a) included only a small sample of languages, or (b) included a large sample of languages but failed to control for diversity of structural, genealogical, or areal factors. The identification or refutation of universals (or superabundant cross-linguistic tendencies) in language acquisition requires that the sample of languages approximate as closely as possible the full diversity of linguistic structures in human language (cf.) Stoll & Bickel 2013). Languages from the same language family are more likely to share linguistic features, as are unrelated languages in contact situations (e.g., within Sprachbunds). Genealogical and/or geographical affinities between languages tend to result in both cultural and typological similarities. This complex of factors could reasonably yield similar patterns of variation sets across languages in an under-diverse sample and so give a false impression of the generality of the phenomenon. This means that we need more languages, but also languages that have been carefully selected to control for genealogical and areal relationships.

Second, prior work has typically relied on the similarity of full utterances or individual words to identify variation sets. However, as Küntay and Slobin (1996) note, differences in the average morphological complexity of words across languages present problems for cross-linguistic comparison. In analytic languages, searching for word repetitions is relatively straightforward. Take English for example. In (1), the verb *see* is repeated five times, but it has the same form each time; only the syntactic environment changes. By contrast, in morphologically complex languages, word forms themselves change more frequently and more radically. With such languages, utterance- or word-level matching might fail to detect repetition at the sub-lexical level. Consider the following example of two successive utterances from a Turkish mother to her 19-month-old daughter (Küntay & Slobin, 2002: 8; repeated elements given in bold):

(2) Bana odandan bi tane bebek **getirebilirmisin**?
'Can you bring me a doll from your room?'

Getir.
'Bring.'

In (2), no two words are repeated, but the concept 'bring' is. However, the first instance of *getir* 'bring' is suffixed with three markers (*−ebilir* MOD, *−mi* YN, and *-sin* 2S)[1] while the second instance has no additional morphemes. The two sentences also differ in the syntactic context, as in (1). On the one hand, these sub-lexical relationships cause problems for algorithms that rely on word or utterance-level similarity metrics for identifying variation sets. On the other hand, if such algorithms had access to the morphological parse, they would surely discover the obvious match for the verb root *getir* 'bring'.

Third, longitudinal studies of variation sets have only considered age as an index of the child's development (e.g., Grigonytė & Björkenstam, 2016). However, age may not always be the best measure of development, particularly when comparing multiple corpora that span different age ranges (Stoll & Gries, 2009). Furthermore, each child follows a different developmental pathway over time. This means that specific ages – as well as age spans – may not correspond to the same periods of development across children, making any inference about global trends difficult. Fortunately, there exist several more targeted possibilities for measuring development, such as the mean length of utterance or lexical diversity. While none of these measures is without its difficulties, employing them all simultaneously allows us to look for convergent developmental trends. If changes in the prevalence of variation sets are replicated across developmental indices, then we would have stronger evidence that these changes are indeed tethered to increasing linguistic aptitude in the child.

[1] MOD = modality marker (in this case, essentially English 'can'); YN = yes/no question marker; 2 = second-person agreement marker. Conventions adapted from the source (Küntay & Slobin, 2002).

Fourth, the amount of repetition possible in any given language depends on the statistical properties of the lexicon (words or morphemes). For example, a lexicon of two words in a corpus of 10,000 tokens will be very repetitive, more so if the relative probability of one word stands at 90%. More importantly, the levels of repetition that have been reported in the literature could have arisen by chance, or could be much less substantial than they appear on the surface, given a smaller deviation from the expectations given the frequency distribution of word types. We propose to handle this issue by measuring how repetitive a text would be if it were randomly generated using the same underlying lexicon and frequency distribution. This value can then be compared to what we observe in the true text to determine (a) whether repetition is more or less than would be expected and (b) a more reasonable picture of the magnitude of the effect.

Fifth and finally, the developmental trends that have been uncovered so far suggest that variation sets are in fact a special feature of adult-child interaction. However, no study to our knowledge has explicitly tested the behavior of other modes of discourse. For example, do variation sets also exist in adult-directed speech (ADS)? If ADS and CSS are equally repetitive, then the primary explanation for the presence of variation sets would be undermined. ADS thus provides a crucial baseline. Moreover, it allows us to (a) judge the magnitude of the prevalence of variation sets relative to a natural standard (rather than random text) and (b) see whether the developmental trends gradually approximate this adult-directed standard, as expected given the available evidence and theory.

The present study seeks to address each of these issues:

- **Sampling**. The languages considered here were selected to maximize cross-linguistic diversity. Specifically, we analyze longitudinal corpora of seven languages from the ACQDIV database (Jancso, Moran, & Stoll, 2020; Moran, Schikowski, Pajović, Hysi, & Stoll, 2016), as well as the Manchester corpus of British English (Theakston, Lieven, Pine, & Rowland, 2001) from the CHILDES database (MacWhinney, 2000). The ACQDIV database contains several languages which have not yet appeared in the quantitative literature on variation sets, as well as new and larger samples of languages which have been studied before. We therefore expand the linguistic coverage of variation sets research while also providing a means for replication.
- **Morphological complexity**. Repetition is measured both at the level of words and of morphological roots (i.e., roots shed of all grammatical markers). Results of the word-level and morpheme-level analyses are then compared across languages to see whether differences in morphological complexity demand different approaches to measuring repetition.
- **Developmental measures**. Besides age, we consider two additional measures of development: mean length of utterance and lexical/morphological diversity. We also derive a measure of the joint information carried by these three variables to get a more general picture of development.
- **Random baseline**. Randomized versions of all of the corpora are generated and the proportion of variation sets re-estimated. These random estimates are then compared against the original estimates to determine whether variation sets are surprising at all, or whether they are simply unavoidable given the nature of the linguistic toolkit provided to speakers. This comparison also serves as a sanity check on the performance of our algorithm. If it cannot distinguish actual speech from random sequences of words, then perhaps the fault lies in our approach and not with the concept of variation sets generally.
- **ADS baseline.** Variation sets are extracted from two ADS corpora for which we have corresponding samples in ACQDIV (English and Chintang). These estimates are then compared to what we find for CSS. Results inform us about whether variation sets are indeed a special feature of CSS or simply a natural part of human discourse.

We proceed as follows. In the next section, we survey prior approaches to the automatic extraction and analysis of variation sets with a focus on methodology. Based on this survey, we outline our own composite, highly flexible algorithm. The algorithm is applied to our sample of eight CSS corpora and two ADS corpora – as well as random versions of each – to derive estimates of the prevalence of repetition. The results are analyzed statistically. CSS data are analyzed longitudinally (based on several markers of the target child's development). ADS data are analyzed cross-sectionally in comparison to binned versions of the CSS data. We discuss the results in light of the existing literature, and point to several directions for future research.

2. Operationalization of variation sets in the literature

Studies that have sought to extract variation sets automatically from corpora (e.g., Brodsky, Waterfall, & Edelman, 2007; Grigonytė & Björkenstam, 2016; Onnis et al., 2008; Waterfall, 2006; Wirén et al., 2016) have used a number of diverse definitions and methods. Waterfall (2006) defines variation sets as sequences of utterances that (i) belong to the same conversational turn and (ii) relate to the same event, given that they share at least one verb or noun (excluding exact repetitions). She analyzes variation sets over time in a longitudinal study of English (12 mother-child dyads with children aged 1;2–2;6).[2] She finds that the proportion of utterances that belong to variation sets decreases from 17% to 12% during the second year of life (1;2 to 2;6). While we cannot speak to the importance of such a modest decrease, it is nevertheless an indicator of the time-evolving nature of variation sets. Using the same definition of a variation set, Onnis et al. (2008) report an almost doubled rate of 27.9% in the Lara corpus (1;9–3;3 years; Rowland, Pine, Lieven, & Theakston, 2005; Rowland & Fletcher, 2006). In an additional analysis, they loosen the criteria for defining variation sets, including any single-word overlap (not just nouns and verbs). This approach yields 58.6% of the utterances as part of variation sets. Further, they find that 34.9% of word types surface in at least one variation set.

Brodsky et al. (2007) use a slightly different definition of variation sets. They define variation sets as sequences of utterances with a lexical overlap of one or more elements in successive pairs of utterances (e.g., first–second, second–third). They allow a maximum of two intervening utterances and they exclude fillers, pronouns, auxiliaries, WH-questions, proper names and a set of function words. With this definition, Brodsky et al. (2007) reanalyze the data used by Waterfall (2006), resulting in 21.5% of utterances being part of variation sets (twelve mother-child dyads with children aged 1;2–2;6). They further analyze 300,000 utterances from the English component of the CHILDES database (MacWhinney, 2000). They find that 18.3% of the words occur in variation sets. The divergent results obtained illustrate that differing definitions of variation sets and how they have been operationalized have an impact on how many variation sets are identified in a particular corpus.

Wirén et al. (2016) define a variation set in a novel way. To identify variation sets in their Swedish corpus of parent-child interactions (Björkenstam & Wirén, 2014), they do stepwise comparisons of successive utterance pairs using Ratcliff–Obershelp pattern recognition (Ratcliff & Metzener, 1988), thereby allowing for maximally two intervening dissimilar utterances within a certain similarity threshold. The Ratcliff–Obershelp algorithm computes the similarity of two strings by matching all characters and then dividing by the sum of the number of characters in the two strings. Matching characters start with the longest shared character subsequence between two strings and then recursively match shared subsequences on either side of it (Ratcliff & Metzener, 1988). First, they identify variations by hand in a corpus of Swedish child-directed speech to create a gold-standard database. They find that variation sets gradually decrease in number as the child gets older. Then they evaluate how well their automatic procedure aligns with the hand-annotated data. The algorithm achieves 0.56 (strict matching) and 0.82 (fuzzy matching) F-scores for the youngest age group, but performance decreases over time. Next, they apply the algorithm to English, Croatian, and Russian, and find that across 4 age groups (0;7–0;9, 1;0–1;2, 1;4–1;7, 2;3–2;9) there is a decrease in the proportion of variation sets in the child-directed speech. The proportion of verbatim repetitions of utterances also decreases.

Grigonytė and Björkenstam (2016) expand the approach by Wirén et al. (2016). They implement a novel method for variation set detection by combining two pairwise comparison strategies, called *anchor* and *incremental*, together with two algorithms for lexical distance comparisons (discussed in detail below): the Ratcliff–Obershelp pattern recognition method and the Python module *difflib*, which is a library that provides string similarity measures, including *edit distance* (Levenshtein, 1966).

An illustration of the anchor method is given in Section 4.1 in Fig. 2. By contrast, incremental comparison simply involves stepwise comparison of successive utterances, e.g., utterances 1–2, 2–3, 3–4, etc. Grigonytė and Björkenstam (2016) compare their results using these two approaches. Their open-source *Varseta* software achieves only low precision and recall figures when applied to the gold standard datasets of Swedish (Wirén et al., 2016) and French (Grigonytė & Björkenstam, 2016).[3] That is, *F*-scores perform relatively poorly across the board – regardless of the anchor or incremental stepwise analysis. The differences between these analyses also depended on the type of matching. The anchor method performed better for fuzzy matching, while the incremental approach performed better for strict matching. However, the anchor method always outperformed the incremental method for the youngest age group (0;6–0;9), and the superior performance of the incremental method for strict matching was negligible up until about two years of age (see Table 4 in Grigonytė & Björkenstam, 2016).

Grigonytė and Björkenstam (2016) then apply *Varseta* to 26 corpora from the CHILDES database (MacWhinney, 2000) with the aim of investigating how the proportion of variation sets changes as children grow older. Given what has been described in the previous literature, they expected to find a decrease for all languages. Although they observe that the proportion of utterances belonging to variation sets indeed decreases for the majority of languages (19 out of 26), two sets of exceptions were identified. The first included Chinese Mandarin, Thai, Hebrew, and Tamil. For these languages, the authors state that there is an insufficient amount of data for the earlier age groups. The lack of data skews the estimates of variation set proportions making it difficult to compare to proportioins across age groups. For example, Grigonytė and Björkenstam (2016) note that the corpus data for children in the Chinese Mandarin corpus in age group 2 (1;0–1;3) contains only 294 utterances; compared with 1395 utterances in age group 3 (1;3–1;11). Second, French and Portuguese corpus data in CHILDES showed no consistent developmental trends.

Finally, Grigonytė and Björkenstam (2016) acknowledge a serious methodological point regarding surface-based approaches to identifying variation sets. In CHILDES, the corpus annotation scheme contains many elements that do not correspond to segmental properties of the input. These include annotation of perceived pause length, prosody, etc. Additionally, certain CHILDES transcripts, such as Cantonese, Chinese Mandarin, and Japanese, are encoded in Latin characters in some cases (about half of the recording sessions), whereas transcripts for the age group 2 are in Chinese (logogram) characters. This difference makes it difficult to compare samples not only within, but across languages. Addressing these challenges when comparing corpus data from heterogeneous sources is critical. As we discuss below, careful attention must be paid to transcription conventions in order to make different corpora

[2] Waterfall (2006) notes that the data used in the study were collected by Goldin-Meadow, Huttenlocher, & Levine (2002–2007) for a project funded by NIH Grant # POl HD40605.

[3] https://github.com/ginta-re/Varseta

syntactically, and semantically, interoperable for cross-linguistic comparison and analysis.

3. Materials and methods

In this section, we describe the data and introduce a comprehensive approach to automatic, cross-linguistic extraction of variation sets. Based on the findings of Grigonytė and Björkenstam (2016), we focus on the anchor method, as it has been shown to perform better against gold standard datasets than the incremental method. Because we are comparing different languages, we further restrict our matches to nouns and verbs. This is because these parts of speech are morpho-syntactically well-defined and productive in all of the languages in our sample.

Section 3.1 describes the data in greater detail. Section 3.2 describes the general form of the algorithm, as well as its parameters. Section 3.3 outlines our additional measures of child development and how they are computed.

3.1. Data

The data for this study can be broken down into two classes: CSS and ADS. We describe each in more detail below.

3.1.1. Child-surrounding speech (CSS) data

Our CSS data come from the ACQDIV database (Jancso et al., 2020; Moran et al., 2016). ACQDIV contains longitudinal corpora of child language acquisition from ten typologically maximally diverse languages (Stoll & Bickel, 2013). These corpora consist of transcribed speech recorded in naturalistic settings. The recordings target specific children between the ages of one and six and cover both their speech and the speech of surrounding children and adults. In the present study, we analyze seven of these corpora: Chintang (Stoll et al., 2012), Inuktitut (Allen, 1996, 2021), Japanese (Miyata, 2012), Russian (Stoll & Meyer, 2008), Sesotho (Demuth, 2015), Turkish (Küntay, Koçbaş, & Taşçı, 2021), and Yucatec Maya (Pfeiler, 2021).[4] For comparability with previous studies, we also added the English Manchester corpus (Theakston et al., 2001).[5]

The corpora in our sample represent a wide range of child-rearing situations. For example, the Japanese corpus contains primarily mother and child dyads (majority child-directed speech), while the Chintang corpus contains interactions between children and multiple individuals, as well as a good deal of speech between adults in the presence of the child. However, we do not have reliable coding for whether speech is directed to the child or someone else (overheard speech) for all utterances. This is why we have chosen to collapse these categories under the label "child-surrounding speech."

Table 1 summarizes the data (bold languages have not appeared in any previous quantitative study of variation sets, i.e., half of the current sample). Session counts reflect the number of independent recordings made across all target children. Utterance counts are based on whatever segments of text were considered to be coherent units of speech by expert transcribers. They correspond roughly to clauses but do not strictly align with any syntactic or interactional (turn-based) unit. Table 2 illustrates the cultural, geographical, genealogical, and demographic diversity of the languages in the sample.[6]

The languages in the ACQDIV sample were selected from five clusters calculated via maximum diversity sampling (Stoll & Bickel, 2013) from typological characteristics encoded in the AUTOTYP database (Bickel et al., 2017) and in the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013). This clustering procedure generates maximal linguistic diversity in regard to typological parameters (for precise definitions of the parameters and values, see Stoll & Bickel, 2013, pg 8), including:

- Verb position (word order)
- Degree of synthesis (verbal and nominal)
- Syncretism
- Presence and nature of agreement and case marking
- Polyexponence and inflectional compactness of categories
- Inflectional classes

In Appendix 1, we provide the typological parameters and feature values for the languages in our sample as given in Stoll and Bickel (2013). The resulting language sample allows us to search for universal processes and mechanisms across typologically diverse languages. Examples 4–11 illustrate the grammatical diversity encountered in this sample for several different grammatical features. For example, word order differs radically between the different ACQDIV languages, e.g., SVO in Russian (3), SOV in Turkish (4), and VOS in Yucatec (5)[7]:

3. Ja ne xoč-u salat!
   1SG.NOM NEG want.IPFV-NPST.1SG.S/A salad.SG.ACC.
   'I don't want salad!' (Stoll & Meyer, 2008; session: A05021006; utterance: 68).

4. Abla çay-ın-ı iç-sin
   sister tea-POSS.3SG-ACC drink-OPT.3SG.S/A.
   'Let sister have her tea.' (Küntay et al., 2021; irem32-02sep03–02–00-16; 1825).

(continued on next page)

Table 1
Summary information about corpora used in this study.

Language	Children	Age range	Sessions	Utterances	Words
Chintang	7	0;7.23–4;4.25	475	160,358	459,187
English	12	1;8.22–3;0.2	804	373,934	1,443,404
Inuktitut	4	2;0.11–3;6.12	75	13,935	22,976
Japanese	7	1;4.3–5;1.23	392	246,091	747,485
Russian	5	1;3.26–6;8.12	449	474,905	1,316,322
Sesotho	3	2;1–4;7	129	23,538	82,923
Turkish	8	0;7.28–3;0.24	373	276,279	936,812
Yucatec	3	1;11.9–3;5.4	233	30,240	91,140

Table 2
The language sample.

Language	Spoken mainly in	Language family	# of speakers	Language status
Chintang	Nepal	Sino-Tibetan	5-6 K	Definitely endangered
English	USA, Canada, Australia, UK, South Africa, New Zealand	Indo-European	360 M	Safe
Inuktitut	Canada	Eskimo-Aleut	30 K	Vulnerable
Japanese	Japan	Japanese	128 M	Safe
Russian	Russia	Indo-European	166 M	Safe
Sesotho	South Africa	Bantu	5.6 M	Safe
Turkish	Turkey	Turkic	71 M	Safe
Yucatec	Mexico	Mayan	706 K	Safe

[4] We do not include Cree (Algonquian) due to its small corpus size (only 12 recording sessions from one child). And we do not include Dënësųłinë (Athabaskan) because this corpus is still being actively collected and annotated.

[5] The Japanese and English corpora were taken from CHILDES (MacWhinney, 2000).

[6] Population figures and language endangerment status are from the Endangered Languages Project (http://endangeredlanguages.com/) and the Ethnologue (https://www.ethnologue.com/) and from sources therein.

[7] Examples given here follow the Leipzig Glossing Rules for interlinear glossing: https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

(*continued*)

5. T-u-náach in-k'ab le Osita-o
PFV-3.A-bite POSS.1SG-hand DET O.-DIST.
'That Osita bit my hand.' (Pfeiler, 2021; SAN-1996-06-14; 181).

Regarding degree of synthesis, languages like English are relatively isolating in their morphology, whereas Chintang is polysynthetic (6):

6. Athom u-patt-a-ŋ-s-a-ŋ-ni-ŋ = kha
before 3A-call-PST-1sP-PRF-PST-1sP-3p = NMLZ.
'They had called me before.' (Stoll et al., 2012; CLDLCh2R02S01b; 415).

Whereas English and Russian are nominative/accusative languages (i.e., the nominaitive marks the only argument of an intransitive clause, S, or the agent in a transitive clause, A, , while the object of a transitive clause, P, is marked by the accusative), Chintang and Inuktitut exhibit ergative/absolutive alignment and use a separate form (the ergative) to mark A arguments, while S and P are marked by the same form (the absolutive). Note that there is a good amount of variability in the specific behavior of ergative/absolutive systems across languages. For instance, even though both Chintang and Inuktitut have an ergative that is used to mark agents in (6a) and (7a), the Chintang ergative also serves (among others) to mark causes (6b), while the Inuktitut ergative is also (again among others) used as a genitive (7b)

6a. U-madum-**ŋa** = ta khur-u-gond-o-ko.
POSS.3SG-aunt-ERG = FOC carry-3[s]P-around-3[s]P-IND.NPST[.3sA].
'His aunt carries her around.' (Stoll et al., 2012; CLDLCh3R03S04; 0496).

6b. Kok-**ŋa** = ta me?-no = kha = lo na.
rice-ERG = FOC be.big-IND.NPST = NMLZ = SURP TOP.
'He's so big because of the rice.' (Stoll et al., 2012; CLDLCh2R04S04; 438).

7a. Ii, nuka-pi-ppit atu-ruma-mmauk.
no younger_same_sex_sibling-DIM-POSS.2SG > 3SG.ERG use-want-CAUS.3SG > 3SG.
'No, (it's because) your sister wants to use it.' (Allen, 1996, 2021; LIZ14WM; 206).

7b. Ataata-ppit kami-alu-alu-ni sanarvat-ti-gia-lau-rit.
father-POSS.2SG > 3SG.ERG boot-big-big-MODALIS.DUAL put-CAUS-INCEP-POL-IMP.2SG.S.
'Put your father's big, big boots somewhere.' (Allen, 1996, 2021; ELI51WM; 593).

Another example of typological diversity in the sample regards verbal morphology. Verbs in Japanese (8) do not agree with any arguments, whereas Russian verbs (9) agree with a nominative S/A argument (*obnima-eš* agrees with *ty*) and Sesotho verbs (10) agree with S or both A and P:

8. Okaa-san ga ue kara kore o Otos-u
mother-HON NOM above ABL PROX ACC drop-NPST.
'Mummy drops this from above.' (Miyata, 2012; session: tom20010518; utterance: 1806).

9. Kak ty mam-u obnima-eš'?
how 2SG.NOM mother-ACC embrace.IPFV-PRS.2SG.
'How do you embrace mummy?' (Stoll & Meyer, 2008; session: A00410909; utterance: 594).

10. Mme o-e-hlatsw-its-e
mother(I) NC·I·S/A-NC.IX.P-wash-PRF-IND.
'Mother washed it.' (Demuth, 2015; session: tiid; utterance: 143).

For more examples and detailed discussion of the typological characteristics of the ACQDIV language sample, refer to the ACQDIV corpus database user manual (Moran, Schikowski, Jung, & Stoll, 2021).

3.1.2. ADS data

Conversational ADS data are only available for two of the languages that also appear in our sample of CSS: English and Chintang. Note that these languages differ along several of our typological variables (see Appendix 1), and as such represent a highly contrastive test pair for typological comparison.

The English data come from the spoken component of the British National Corpus 2014 (Love, Dembry, Hardie, Brezina, & McEnery, 2017). This corpus contains roughly 11.5 million words of transcribed conversation. The Chintang data come from a large audiovisual corpus of adult conversation (a subset of Bickel et al., 2011). We restrict the analysis to conversations that did not involve any experimental manipulation (e.g., we omitted conversational retelling tasks, such as those collected using the Pear Film stimulus; Chafe, 1980). The resulting sample consisted of 17 different sessions, totalling 7836 total utterances.

3.2. Procedure

We propose a general procedure for automatically extracting variation sets from text-based corpora. Our procedure is flexible and can accommodate any crossing of the parameters we adopt from previous studies (e.g., Brodsky et al., 2007; Grigonytė & Björkenstam, 2016; Wirén et al., 2016). In this way, we can compare the performance of various combinations of parameter settings across different languages and target children's age ranges.[8]

The ***object of matching parameter*** dictates at what linguistic level the analysis is performed: words or morphemes. Following Waterfall (2006), we restrict the analyses of words and morphemes to nouns and verbs. While including all parts of speech increases the proportion of variation sets that are identified (Onnis et al., 2008), this approach introduces potentially less relevant aspects of repetition (e.g., function words, interjections, and so on). Further, nouns and verbs are ubiquitous across the languages in our sample, while other categories may or may not be shared (e.g., Russian lacks articles, and adjectival meanings tend to be encoded by verbs in Inuktitut, but not in many of the other languages in our sample). Fig. 1 compares two consecutive utterances taken from a corpus of Chintang (Bickel et al., 2011). The first utterance contains a noun root (*thaũ* 'place') and verb root (*ims* 'sleep'). Both words are morphologically complex, carrying the plural suffix *-ce* and the nominalizing suffix *-kha,* respectively. The second utterance contains two noun roots (both *thaũ* 'place') and two verb roots (*yug* 'stay' and *ca* 'eat'). Similar to the first utterance, both verb roots are suffixed with the nominalizer *-kha,* and one noun root is suffixed with the plural *-ce*. These differences in morphological realization lead to different behavior in the word-level and morpheme-level analyses. The morpheme-level analysis produces consistent results as it is blind to the variable morphological contexts. The word-level analysis yields more variable results because it is sensitive to parts of the word beyond the root. This sensitivity depends on how the units are matched, a point to which we now turn.

The ***type of matching algorithm parameter*** includes two match conditions: strict and fuzzy. The strict condition requires exact matches of words or morphemes (i.e., identical strings). Fuzzy matching allows strings to be counted as matches even if they are not identical. Therefore a question is, how similar do two strings need to be? We measure the degree of similarity using the *SequenceMatcher* function from the *difflib* library in the Python programming language, which implements a version of the Ratcliff-Obershelp gestalt pattern matcher (Ratcliff & Metzener, 1988).[9] *SequenceMatcher* returns a similarity score between 0 (no overlap between two strings) to 1 (identical strings). Following Grigonytė and Björkenstam (2016), we set the threshold for a successful

[8] We originally considered an additional parameter, namely, the number of matches required between two utterances before they are considered to belong to a variation set. In a preliminary analysis, we found that even moving from a threshold of one match to two virtually eliminates our ability to detect variation sets. While this fact is interesting in itself, a full investigation is beyond the scope of the present study.

[9] https://docs.python.org/3.6/library/difflib.html

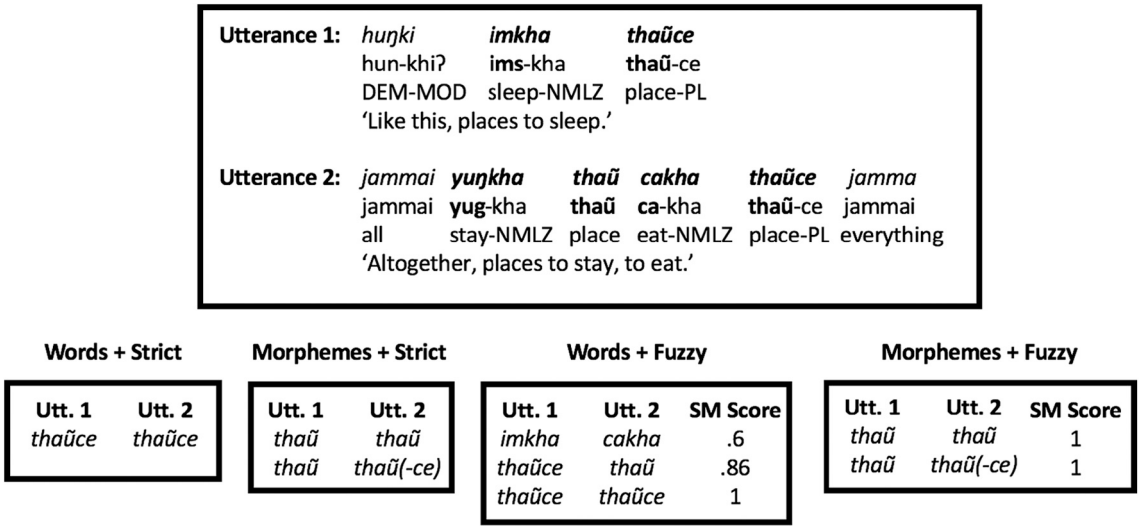


Fig. 1. Different variation set match outcomes for strict and fuzzy matching for words and morphemes in two Chintang utterances. Bolded elements are those considered for matches. Only successful matches are shown. SM score = SequenceMatcher rating of similarity.

match between words or morphemes to 0.55. Fig. 1 exemplifies the differences between strict and fuzzy matching. For fuzzy matching, the similarity ratio is provided. As this example shows, fuzzy matching returns morphologically similar words as a match even when the roots differ (given sufficient similarity of the additional morphemes, in this case, the shared *-kha* in *imkha* and *chaka*).

The ***window size parameter*** defines the number of consecutive utterances that we consider when making pairwise comparisons. In Fig. 1, the window size is set to two (i.e., we have two consecutive utterances), which is the minimum value. For any window size setting *w*, the number of possible pairwise comparisons is *w*-1 (i.e., one comparison in Fig. 1).

To summarize, the parameters in our algorithm include:

- Object of matching (nouns/verbs only; full words or morphemes)
- Type of matching algorithm (exact or fuzzy string matching)
- Window size (number of successive utterances in which we look, e.g., 2–10)

In the remainder of this study, we will consider several different combinations of parameter settings. Specifically, we cross our two objects of match (word and morpheme) and two types of match (fuzzy and strict) to yield four conditions: words with strict matching, words with fuzzy matching, morphemes with strict matching, and morphemes with fuzzy matching. In each of these conditions, we estimate the proportion of utterances belonging to variation sets for windows of size from 2 to 10. While prior research has explored both incremental and anchor-based approaches, we only pursue the anchor method here. This decision was made because Grigonytė and Björkenstam (2016) report better or equivalent performance of this method in all but the oldest age group when it was evaluated against their gold standard. The anchor method is illustrated in Fig. 2.

The anchor strategy measures pairwise similarity of the first utterance in a sequence (the anchor) to each subsequent utterance within the window. The similarity of the words or morphemes in each pair of utterances is then computed. If the number of matching words or morphemes meets or exceeds the required number of matches, both utterances are labeled as belonging to a variation set. In this study, we only consider a match threshold of one. This way, we maximize the number of variation sets that we observe. We then iterate the process by sliding the window forward through the text, one utterance at a time. Crucially, the status of whether any utterance belongs to a variation set can only move from *no* to *yes* across iterations. That is, an utterance must only match one other utterance in one window to be counted as belonging to a variation set, and that status is final.

The outcome of each of our analyses is a proportion representing the number of utterances that belong to at least one variation set out of the total number of utterances (0 = no utterance belongs to a variation set; 1 = every utterance belongs to a variation set). We compute one such proportion per session, per corpus, per parameter combination.

Finally, we repeat the steps outlined above on a simulated randomized version of each session, for each corpus, and for both words and morphemes (where available). To create the randomized versions of the texts, we first calculated the average number of nouns and/or verbs per utterance, per session. This average amounts to the effective mean length of utterance (eMLU) since our matching algorithm only compares nouns and verbs. Next, for each session, we generated *n* random utterances of length *l*, where *n* is equal to the number of utterances in the original session and *l* is equal to the eMLU of that session. Each random utterance was constructed by sampling with replacement from the list of all noun and verb forms (words or morphemes, depending on the level of analysis) that appear in the target session (tokens, not types). The result is a random session which consists of the same number of utterances as the original session, where each random utterance is (1) as long as would be expected based on the eMLU of the original and (2) sampled according to the same empirical distribution of word forms as is present in the original.

Random baselines defined in this way have certain properties. Most importantly, they may produce estimates of the amount of repetition that are greater or less than those observed in the corresponding true sample. If greater, there are two potential explanations. First, it could be that there really is a superabundance of repetition in the true sample. Such a situation would be fully in line with the traditional understanding of variation sets. It would mean that even though a few tokens take the lion's share of the probability mass, when the others do appear, they do so in clusters. Second, the distribution from which the random sample is computed might not be particularly Zipfian in the first place. If it shows a fatter positive skew (many more higher frequency types), then simple random matching becomes more difficult. But note that such a fat-tailed distribution would be informative for understanding CSS: it would suggest that adults use richer, but more clustered vocabulary when speaking with children, in line with the first point (more repetition, but clustered) and variation set theory overall. If the true estimates are lower than the random estimates, then there is only one explanation: speakers are systematically less repetitive than they could have been. In

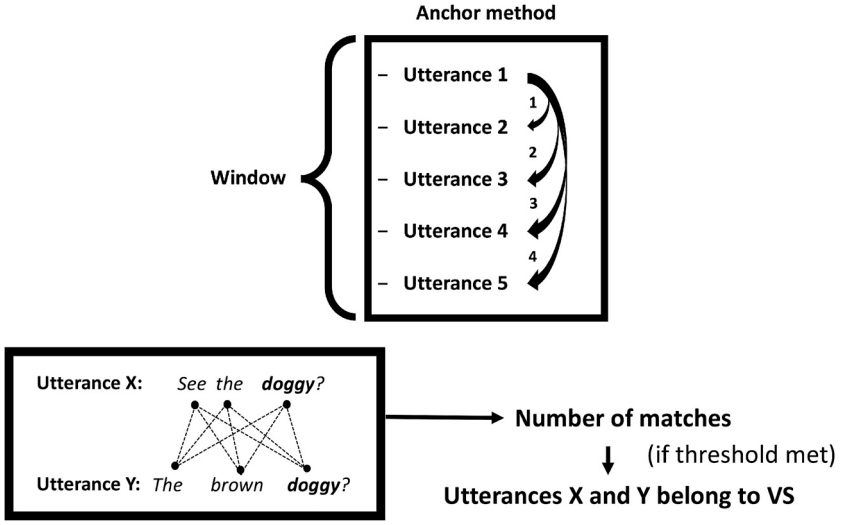


Fig. 2. Anchor method and how comparisons are made between utterances (for an example window size of 5). Arrows in the top panel indicate successive comparisons against the first utterance. Dotted lines connect words (or morphemes) that will be compared between utterances.

other words, they prefer to be informative, and not simply produce words in accordance with their frequency distribution. In this case, the difference between the random and true estimates becomes particularly important. The closer they are, the less informative the speech is. It communicates the frequency distribution of words but without the special degree of repetition that would suggest the presence of variation sets. The more distant the two estimates, the more the speakers have gone out of their way to avoid the Zipfian curse on repetition in their local choices. They become more informative, and less repetitive, relative to their statistical potential. This difference is therefore a more fine-tuned instrument for gauging the presence and/or degree of variation sets in a given text than the simple empirical proportion itself.

3.3. Measures of child development

To address the potential issues of treating age as an index of linguistic development, we consider several additional indices derived from the speech of the children: mean length of utterance in words (MLUw), mean length of utterance in morphemes (MLUm), lexical diversity (lexical H) and morphological diversity (morphological H).

Mean length of utterance is simply the total length of a given session in words or morphemes divided by the total number of utterances in that session. We compute these measures for all sessions where possible (some sessions do not have morphological annotation).

Diversity is operationalized as the Shannon entropy of the frequency distribution of words or morphemes (bias-corrected using the method described in Chao, Wang, and Jost (2013). As with the MLUs, entropies were computed for each session. Each entropy estimate was then normalized by dividing it by the associated maximum entropy for that session (defined as the log of the number of distinct word or morpheme forms in the session).

Age, MLUm/w and lexical/morphological H are all tightly intercorrelated. We therefore attempted to identify their common information by means of a principal component analysis (PCA). PCA produces n orthogonal rotations of a matrix of variables (components), where n is the number of columns in the matrix. These components can then be compared individually against the original matrix to give some idea of what they encode. We performed one PCA over age (log transformed), MLU, and H for each language in our CSS sample. Prior to each PCA, all variables were centered using z-scores. To keep level of analysis consistent, where word-level and morphological data were available for a given language, we performed two PCAs: one with age, MLUw, and lexical H; and one with age, MLUm, and morphological H. In all languages, the principal component (i.e., the component that explains the most variance) showed at least two of the three developmental variables in alignment. In the cases where only two variables aligned (Inuktitut–morphemes and Sesotho–words), the third was either non-correlated or marginally negatively correlated. A typical example is given in Fig. 3.

In Fig. 3, the x-axis plots the principal component (PC1); the y-axis plots the second-most informative component (PC2). The portion of cumulative explained variance associated with each component is given in parentheses on the axis label. Arrows indicate the direction and strength of association of each to the principal component, and each is labeled for the original variable it represents. Henceforth we focus only on PC1. Right-facing arrows indicate positive association with PC1. Deviation from a slope of zero, either up or down, indicates decreasing association strength. In this case, all arrows point in the same direction with respect to PC1, indicating that this component captures information shared across the developmental indices. PC1 is therefore a more distilled representation of overall linguistic development than any of the source variables individually, albeit one that more directly corresponds to lexical diversity (WordEnt) and age (logAge) than to MLU (MLUw).[10]

Finally, we bin the CSS sessions based on each of the developmental indices mentioned above to allow for cross-sectional comparison with ADS. Bins were defined so that the total population – across all corpora – was split evenly into four groups so that each bin contains approximately one fourth of all data points from the entire sample of languages. Bin labels run from 1 (lowest values, i.e., earliest stages of development) to 4 (highest values, i.e., latest stages of development). Note that because the corpora vary in the number of children and age ranges covered, not all languages have data points in all bins. The benefit of this approach is that we see directly where one language may be compared to others. The ADS corpora are assigned to a single group for all indices labeled "adult". The result is six new variables: *age groups*, *MLUw groups*, *MLUm groups*, *lexical diversity groups*, *morphological diversity groups*, and *PCA groups*.

4. Results

Here we report the results of two sets of analyses, one longitudinal and the other cross-sectional. The longitudinal analyses look at how

[10] All PCAs, as well as visualizations of their results can be found in our GitHub repo (https://github.com/acqdiv/variation-sets).

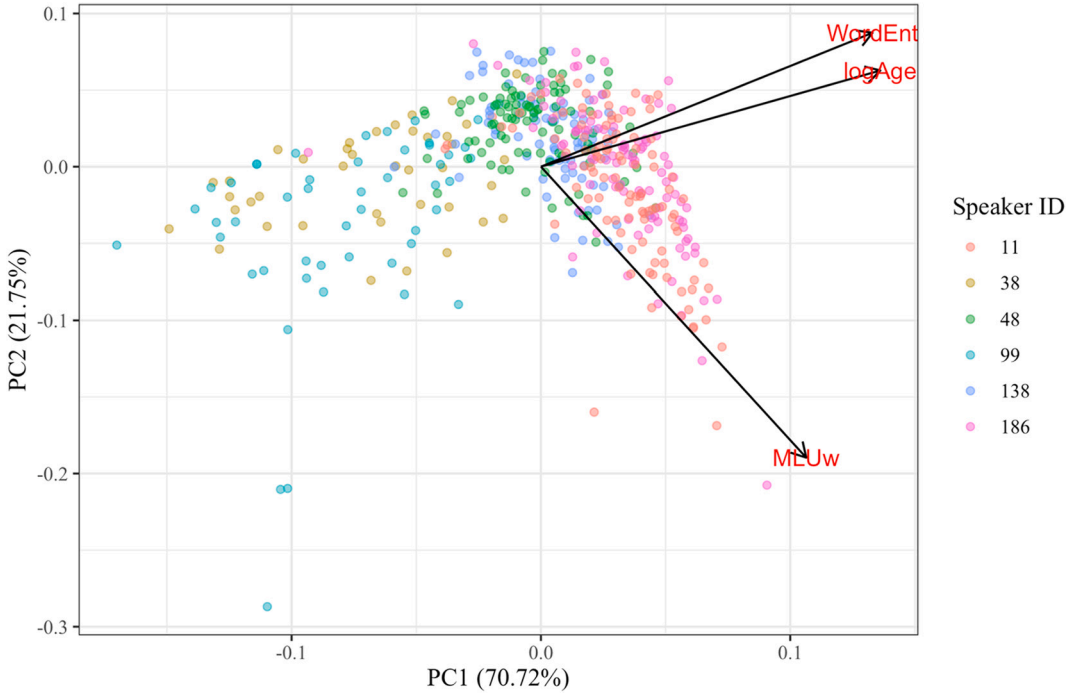


Fig. 3. PCA results for the word-level analysis in Chintang.

repetitiveness relates to the development of the target child, measured in several different ways. The cross-sectional analysis compares the average repetitiveness in CSS at different developmental stages to the average repetitiveness in ADS. All analyses are compared across true and randomized versions of the samples.

4.1. Longitudinal analysis of child-directed/surrounding speech

Due to the complexity of the parameter space, we perform several linear mixed-effect regressions using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) from *R* (R Core Team, 2021). Each regression was based on a different dataset, and each dataset captured a unique crossing of match type (two levels: *words* or *morphemes*) and match criterion (two levels: *identity* or *fuzzy*, where a fuzzy match requires a SequenceMatcher score of 0.55 or greater as per Grigonytė & Björkenstam, 2016; see Section 3.2). This process resulted in four general types of models:

- words & fuzzy
- words & strict
- morphemes & fuzzy
- morphemes & strict

Russian, English, and Japanese were omitted from the morpheme models because the words have not been morphologically segmented.

The type of model was further crossed with the type of developmental index – age, MLU, lexical/morphological diversity, and the principal component taken over all three – to produce sixteen total individual model types. Variation sets were modeled in each language separately using as many of these model types as possible given the limits of the data.[11]

All of the models share a common structure. The response variable was the estimated proportion of variation sets from each recording session. Because proportions are inherently bounded between 0 and 1, we apply a logit transform before fitting the regression. This transformation stretches the range of possible values to $\{-\infty, \infty\}$, which prevents the model from generating expected values outside of the possible range. Fixed effects are:

- developmental index – age, MLU, word/morpheme diversity, or PCA component based on the target child
- window size
- number of utterances in the session
- number of unique speakers in the session
- mean length of utterance (MLU-adult; nouns and/or verbs only)
- text type (*random, original*)

We further include the two-way interaction between text type and developmental index. Random intercept adjustments were added for recording session nested in target child (each session is uniquely associated with one target child, but each target child appears in many sessions).[12] To increase the reliability of our estimates, we restrict the analysis to sessions with 50 or more total utterances. We also remove outliers, defined as observations of the dependent variable that fall outside the range of two standard deviations above or below the mean per target child within each crossing of corpus, level of analysis, and type of match.[13]

[11] We also performed a model over the entire sample with a corpus x developmental index interaction term. All models produced similar results. See the https://github.com/acqdiv/variation-sets/blob/master/Analysis/analysis.Rmd for the complete set of models and https://github.com/acqdiv/variation-sets/tree/master/Results for all model summaries.

[12] The models for Inuktitut would not converge with the full random effect structure, so we simplified to random intercepts for the target children. The same was true for Japanese (fuzzy, morphemes) and Yucatec (strict, morphemes, age/PCA/MLU). These changes only minimally affected the model estimates.

[13] We also perform the same models for the full samples (i.e., with no outliers removed), as well as samples for which outliers are defined by the interquartile range. The approach we report here leads to the best behaved models. All of the models are reported in https://github.com/acqdiv/variation-sets/blob/master/Analysis/analysis.Rmd.

For each model, we take a two-step approach. First, we test the true data against the developmental index (with all additional controls) to determine whether there is a significant association. Then, we add the random data, along with the main effect of text type and its interaction with the developmental index. In this way, we attempt to determine whether significant main effects of the index can be distinguished from chance. The complete set of models and model summaries, along with all necessary data, can be found in the GitHub repository.

4.1.1. Word-level analyses

We begin with the control variables. The only control to be significant in all languages, in all conditions, was window size. Increasing the window size always increases the proportion of variation sets. This makes perfect sense: adding utterances increases the number of possible comparisons, and with each new comparison comes another chance for a match. Number of speakers was significant about 60% of the time, again always in the same direction. Adding speakers decreases the observed proportion of variation sets. MLU-adult was significant about about 40% of the time and was also consistent: increasing numbers of nouns and verbs per average token leads to higher match proportions. Finally, the number of utterances was nearly never significant at only 12%. When it was, the correlation with variation set proportions was positive. Longer sessions occasionally yield higher proportions of variation sets. Please see the GitHub repo for the complete set of model summaries.

The critical variables are summarized in Figs. 4–5. In each figure, columns correspond to developmental indices and rows to corpora. In each panel, y-axes correspond to the expected proportion of utterances belonging to a variation set, and x-axes correspond to the developmental index (centered with *z*-scores). Regression fits are plotted as lines. The 95% confidence intervals are plotted as colored ribbons around the regression lines. Stars appear immediately adjacent to the developmental index label in the title of each plot to indicate the significance of the interaction with text type (***$p \leq .001$, **$p \leq .01$, *$p \leq .05$; these conventions are followed throughout). If no stars appear, the relationship is not significant (i.e., the random baseline and the original text behave the same over development). Stars also appear within parentheses next to the plot title. These indicate the significance of the effect of the developmental index when tested in isolation (i.e., whether there is a reliable association between the developmental index and proportion of variation sets in the true sample alone). The star system works the same, but we include *n.s.* to indicate non-significance. We begin with the word-based analyses.

Looking first at the simple analyses (i.e., those without the random data; significance indicated in parentheses above each panel in Figs. 4–5), we see that every language except one showed a significant correlation between at least one developmental index and the proportion of variation sets, regardless of whether we apply strict or fuzzy matching (the one outlier of the group is Inuktitut, for which no significant correlations were uncovered). Note also that there is remarkable consistency in the shape of the effects across developmental indices within each language. However, there are cases in which not all indices reach significance, and these gaps are not consistent across languages. Therefore, our decision to include as many such indices as possible is justified not only because it allows for the discovery of convergent evidence, but also because it corrects for differences in the behavior of any single measure across corpora (based on their contents and/or structure) or languages.

Variation sets were detected for all languages in all conditions (i.e., in no language was the proportion universally zero). We therefore replicate the findings of prior cross-linguistic studies (e.g., Grigonytė & Björkenstam, 2016). However, there is considerable variation in at least two respects: the range of proportions predicted and the direction of the developmental effect. The range of proportions is largest for Russian, which reaches as high as 50% and as low as 5%. Turkish also shows a large range, covering values from about 10% to 40%. Sesotho shows the weakest effect, ranging only between approximately 2 to 10%. Regarding the direction of the developmental effect, we see a three-way split. English, Japanese, Russian, Sesotho, and Yucatec all show the expected downward trend over time. English, Japanese, Russian, and Sesotho show the effect in at least three out of four developmental indices for both fuzzy and strict matching. Yucatec only shows the effect for age in the fuzzy matching condition, but shows it universally for strict matching. Inuktitut showed no developmental effect for words in any condition. But surprisingly, Chintang and Turkish (and numerically, Inuktitut, in some cases) show increasing amounts of repetitiveness for at least one of the four indices. Age was always significant, and Turkish showed significant effects for all indices in all conditions. Again, the effect sizes differ, with Turkish showing a much stronger trend than Chintang. This finding is particularly surprising given that prior studies have reported decreasing trends for Turkish using age as the developmental index (but based on a different corpus; Grigonytė & Björkenstam, 2016). Nevertheless, the fact that at least two languages in our sample exhibit increasing amounts of repetition, and that this effect was replicated for at least two developmental indices in each, strongly suggests that the reversal in the trend is not a fluke.

Turning to the interaction effect, it was significant for all languages for at least one developmental index, regardless of matching method. This means that there is some structure to the developmental curves that we see for the true texts which cannot simply be attributed to random selection from a vocabulary. However, as implied above, we do not detect a difference for all languages in all conditions, nor for all developmental indices. This fact underscores the importance of exploring as many developmental indices as possible when comparing variation sets across languages.

In the fuzzy matching condition, random estimates were uniformly greater than the true estimates. Thus, the corpora were less repetitive than would have been expected by chance. The same general pattern holds, though to a much lesser extent, in the strict matching condition. For the majority of languages, the random and true estimates are much more similar to each other than was observed in the fuzzy-matching condition. This difference appears to be driven by overall decreases in proportions observed between the fuzzy and strict conditions. When the proportions for the true text, even based on fuzzy matching, were low, the overall decrease in proportions due to strict matching leads to a closer approximation between the true data and the random baseline. Three languages buck this trend. Sesotho and Inuktitut (for the most part) maintain a larger gap between the random and true estimates. Russian shows the opposite pattern, with true estimates topping random estimates at most developmental ranges, though the estimates remain quite close to each other.

The greater overall similarity between random and true texts when strict matching is applied indicates that the strict-matching criterion reduces our ability to distinguish signal from noise in variation set proportions, even when the differences can still be established as reliable. But the major take-away is that different corpora – either because of their content or the intrinsic properties of the languages – respond differently to randomization with respect to the behavior of our matching algorithms. We cannot simply assume that the proportions we see in true texts across languages are indeed comparable. The more they deviate from the random baseline, the more information is carried by the repetition. Hence, the sheer magnitude of the proportions themselves is potentially misleading.

But what of the shape of the random baseline curves? While most languages show a generally flat curve (given in red in Figs. 4–5), this is not universally the case. For example, Turkish words with strict matching show a decreasing random baseline over time. These small trends are most likely due to shifts in the underlying frequency distribution of words across the corpora. If the samples become less broadly repetitive (i.e., more Zipfian), over development, then we should expect the random baseline for variation set proportions to decrease. In other words, time-evolving random baselines suggest global shifts in the frequency distribution of words. Notably, the random baselines in Inuktitut

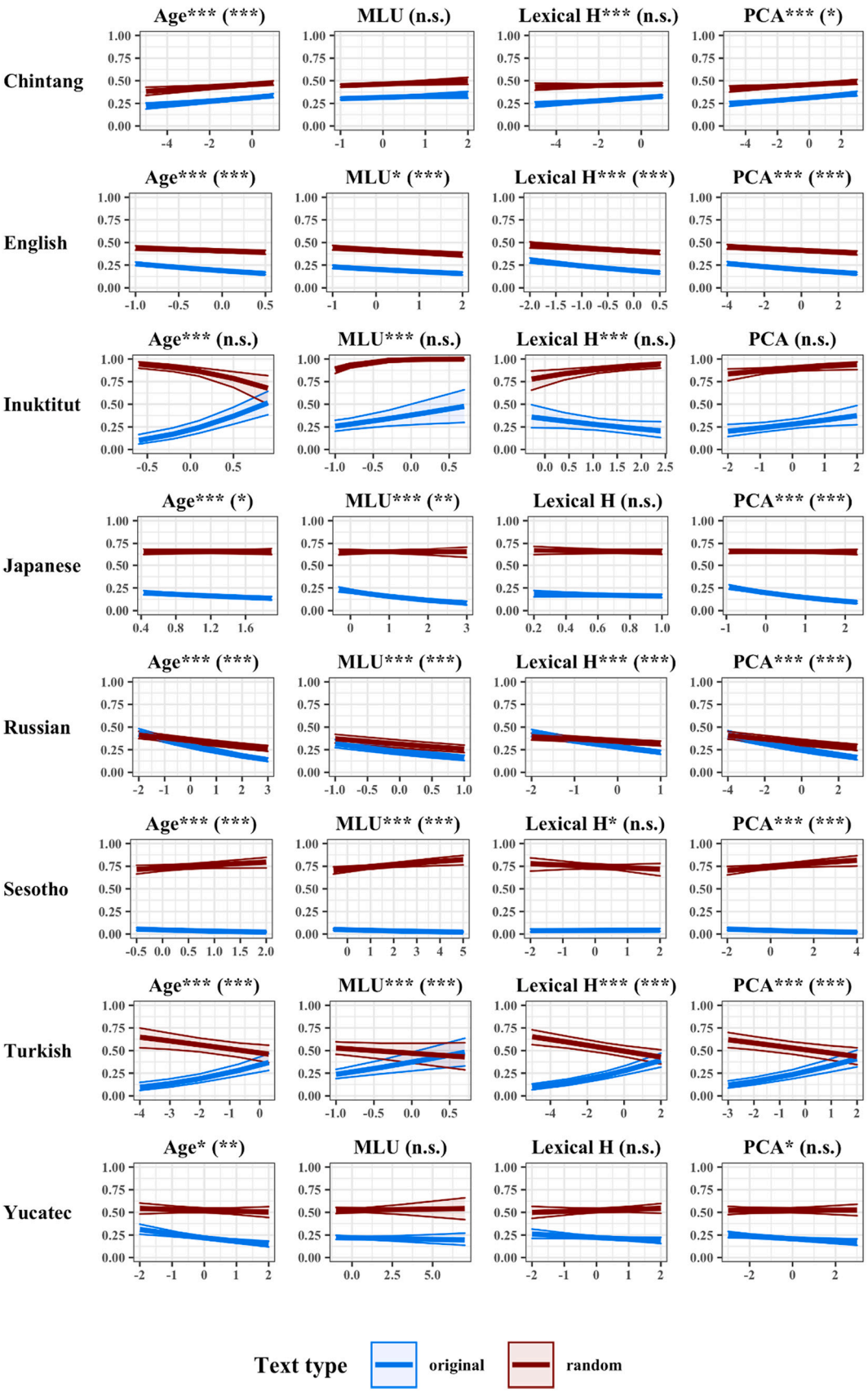


Fig. 4. CSS, words, fuzzy matches.

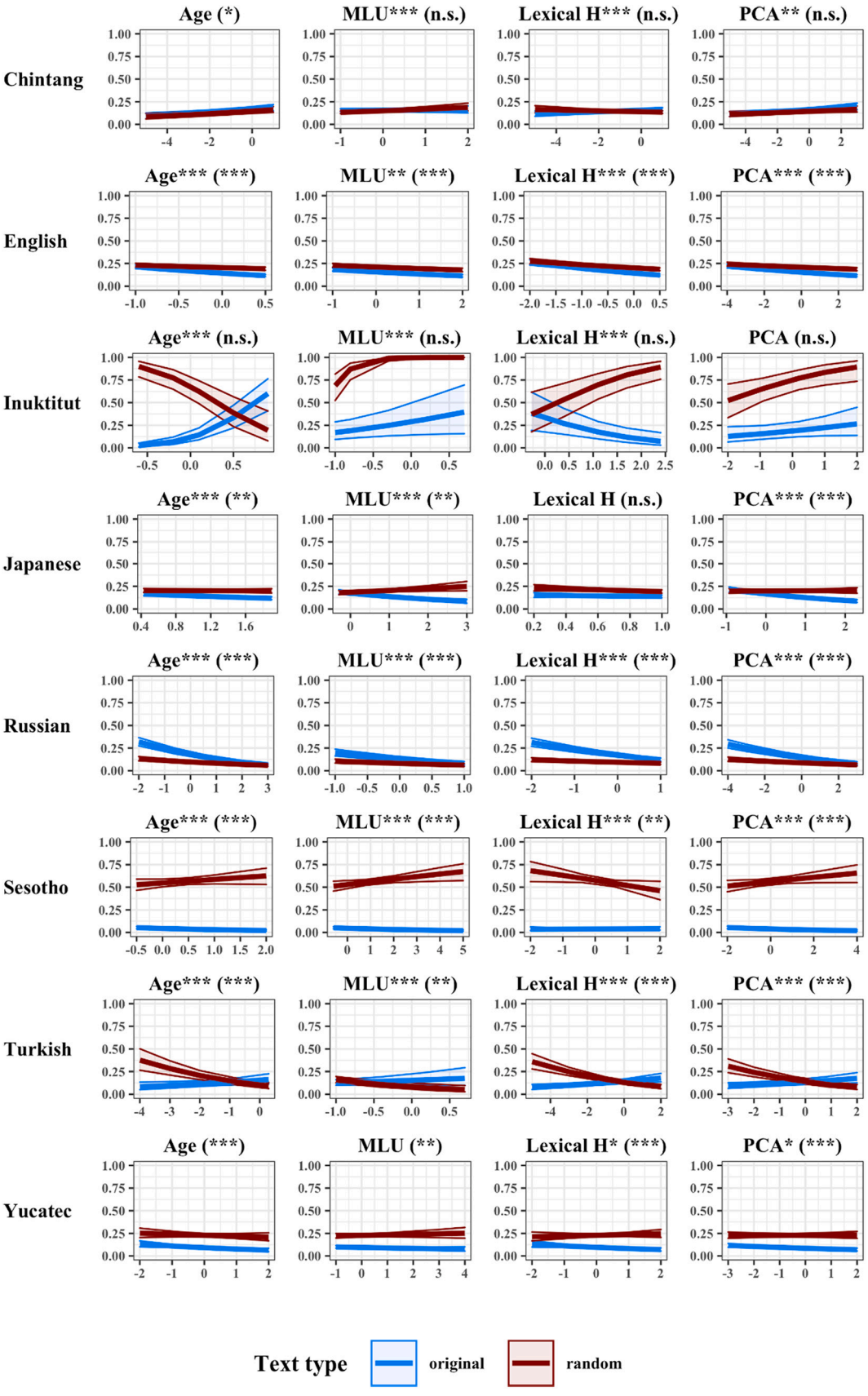


Fig. 5. CSS, words, strict matches.

behave somewhat erratically, even changing direction across the different developmental indices. The source of this behavior is unknown. One contributing factor could be the fact that this corpus is much smaller than the majority included here. For example, it is only one-quarter the size of the next largest corpus, and less than 4% of the median size across corpora. Small samples lead to less accurate estimates of linguistic behavior, especially MLU and lexical entropy (hence the PCA), which in turn could produce erratic patterns of development, even in the randomized version of the corpus. This fact might also help to explain why we did not uncover any reliable developmental effects for this language.

4.1.2. Morpheme-level analysis

We repeat the above analyses for the morpheme-level data. Recall that only five of the eight total languages were annotated for morphological structure, and as such, suitable for this analysis. Note also that this subsample includes the languages with most elaborate morphological structures, particularly in the verbs (Chintang, Inuktitut, Turkish, Yucatec). Results are presented in Figs. 6–7 for fuzzy and strict matching, respectively. Plotting conventions are the same as those in Figs. 4–5.

The overall pattern of results for morphemes is consistent with that observed for words. The major difference is that the morpheme-level analysis yields greater proportions of variation sets overall. In some cases, developmental trends were strengthened, especially for Turkish, and Yucatec. As with words, changing morpheme-level matches from fuzzy to strict reduces the proportion of observed variation sets in all languages. We also see the now familiar trend for true data to more closely approximate the random baseline with strict than with fuzzy matching. For all languages in the sample, at least two of the developmental indices differ significantly in their developmental trajectory from the random baseline, with the exception of Sesotho in the fuzzy matching condition.

Sesotho is an outlier in the group for another reason; it shows no reliable developmental trends when considered at the level of morphemes. This contrasts with the word-level analysis, in which reliable, if somewhat weak, negative trends were observed in both fuzzy and strict matching conditions. However, there are two cases in which Sesotho shows a significant interaction between developmental index and text type (morphological diversity in fuzzy matching; MLUm and PCA in strict matching). In all of these cases, the random baseline shows a slight positive trend, which leads to a greater difference between it and the true estimates at the latest developmental stages. As discussed above,

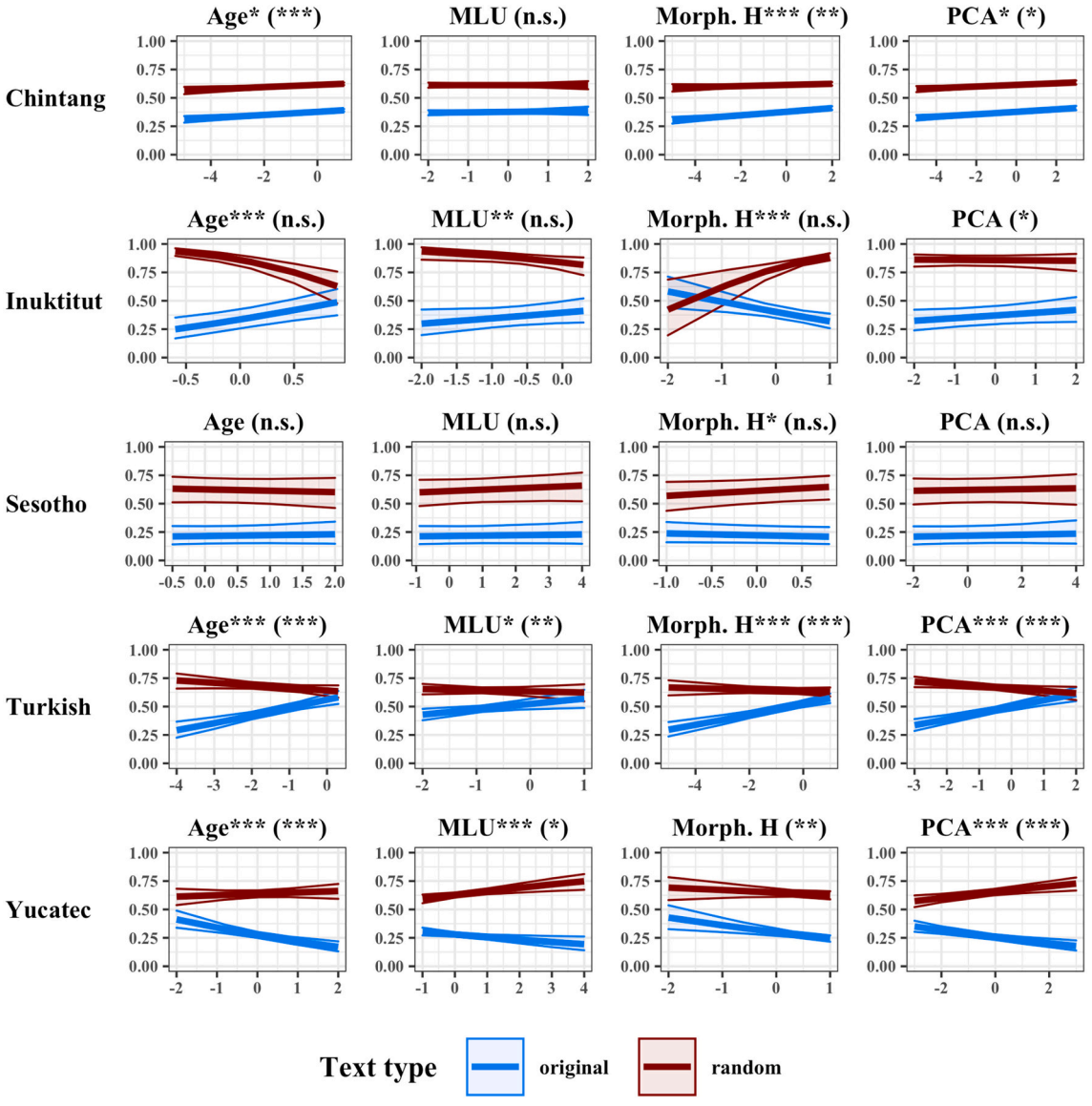


Fig. 6. CSS, morphemes, fuzzy matching.

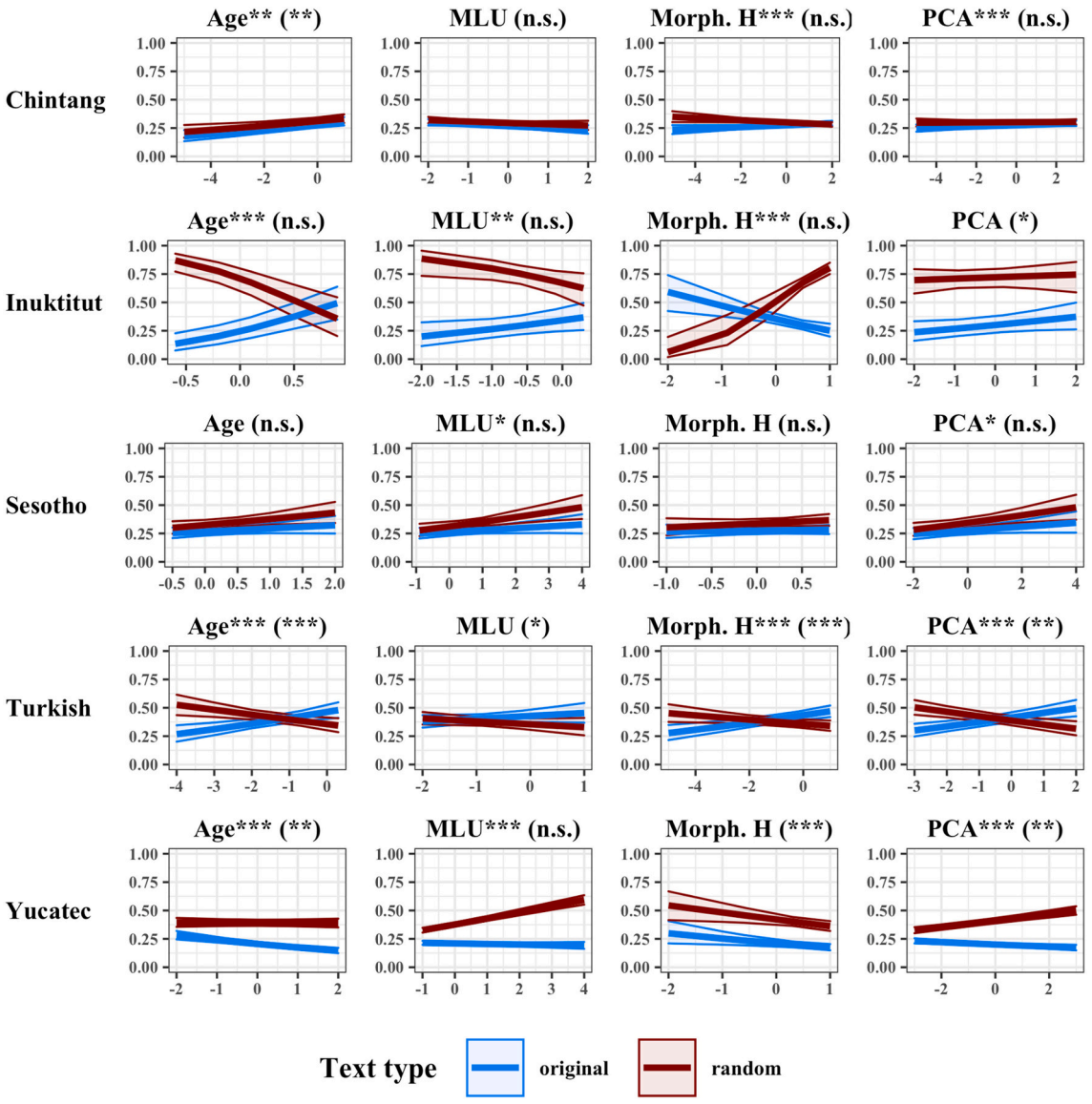


Fig. 7. CSS, morphemes, strict matching.

this difference reflects increasing informativity of the text relative to its underlying frequency distribution. Sesotho thus becomes increasingly less repetitive than it could be over time, a pattern that is consistent with the reliably negative developmental trends observed for words.

On the other hand, Inuktitut showed its only significant developmental trends in the morpheme-level analysis, both of which appeared for the PCA-derived developmental index. But this change is almost certainly a reflex of shifts in the underlying distribution of morphemes, as evidenced by the lack of an interaction effect. Inuktitut only becomes more repetitive because the distribution becomes more sharply Zipfian (i.e., as the probability mass becomes increasingly bunched into a smaller set of types).

The other languages behave more or less as expected given the word-level analysis. This indicates that for many languages – even some that are morphologically complex – words are a good-enough proxy for detecting variation sets and their longitudinal trajectories. This fact is useful, as morphological analysis of large scale corpora is costly, both in time and effort. Importantly, however, the magnitudes of the estimates and the steepness of the trends depend on a combination of the language and level of analysis. Furthermore, choice of one level of analysis over the other can obscure otherwise present effects (as was the case for Sesotho).

4.1.3. *Interim summary*

Our analysis so far reveals that variation sets appear in some measure in all languages, at all age ranges, irrespective of whether we consider words or morphemes, or whether we require strict matches or fuzzy matches. However, the proportion of utterances that belong to variation sets differs across languages and conditions, as does the relationship between the true and randomly generated text.

The overwhelming trend is for true text to be less repetitive than the randomly generated text. This pattern is much weaker in the strict matching condition for some languages, most notably Chintang and Turkish. Nevertheless, both languages deviate significantly in their longitudinal trends from the random baseline. The tendency for random speech to be more repetitive most likely reflects the informativity of true speech. Randomly sampling from a Zipfian distribution (few high frequency types with a long tail of singletons) results in resampling of the same words, and hence is more likely to produce a match. The difference between this random baseline and the true text tells us something about how much information is carried by actual speech assuming a fixed lexicon and frequency distribution. When the gap between the two types

of text narrows, the repetitiveness detected by our algorithm is more likely to arise simply because of the structure of the lexicon in use (i.e., the words and their relative probabilities), and so theoretically irrelevant to the study of variation sets.

Testing against the random baseline also revealed that an otherwise significant developmental trend in Inuktitut (for the PCA index with morphemes) was not reliable. When these random and true trajectories cannot be distinguished, we must assume that both are driven by the same underlying shifts in the frequency distributions of the target unit (words or morphemes). Future research should therefore pay careful attention to global statistical features of the text in order to safely interpret apparent developmental trends. Even so, the vast majority of languages and conditions did show trends that were significantly different from what would have been expected by chance. This point validates the more general hypothesis that changes in the degree of repetition in CSS are tied to the child's linguistic development.

Regarding the shape of the developmental trends, most languages showed the expected negative or null association. However, two, or possibly three, languages bucked the trend: Chintang, Turkish, and (possibly) Inuktitut. For these languages, CSS tends to become *more* repetitive over time. This effect is preserved even when considering strict matching of morphological roots, which represents the most conservative level of repetition. For Turkish, this finding is at odds with what has been reported previously. Grigonytė and Björkenstam (2016) found decreasing trends of variation sets, though their analysis differed in several ways, including which utterances were considered for comparison (step-wise comparison of adjacent utterances) and how they were compared (global similarity of full utterance strings).

4.2. Cross-sectional analysis of adult-directed and child-surrounding speech

We have so far observed reliable developmental trends in all languages, and these trends were not entirely attributable to the statistical properties of the corpora. However, we have not yet addressed another crucial component of variation sets, namely, their specificity to interactions between adults and young children. Based on prior research, we expected variation sets to become less prevalent as the child becomes a more proficient speaker. By extension, adult-directed speech should be the least repetitive. By comparing CSS and ADS directly, we can test this hypothesis. We can also assess the amount of repetitiveness that distinguishes adult-directed and child-surrounding speech.

Proportions of variation sets were extracted from true and randomized versions of the spoken BNC and Chintang conversational ADS corpora. The CSS data were split into four evenly sized groups based on each of the developmental indices (i.e., four sets of four groups). Both the English and Chintang data were considered simultaneously when making the splits. This means that there are some gaps in the group coverage depending on the language and the developmental index. For example, the Chintang CSS corpus includes children older than those found in the English CSS corpus, so that for English we only have observations for groups 1, 2, and 3 in the age index. All ADS data is lumped into a fifth "adult" group.

We again model the data using linear mixed-effect regression. The developmental groups were each modeled separately for each language and match type (16 total models; 4 indices $\times$ 2 languages $\times$ 2 match types). As with the prior analysis, we only include sessions with 50 or more utterances in the interest of maximizing the likelihood of observing variation sets. Only word-level matching is implemented as we have no morphological parse of the English data. Prior to modeling, we remove outliers as in the longitudinal analysis (observations of the dependent variable that fall two standard deviations above or below the mean).

The critical predictor in each model is the developmental grouping factor. As before, we also include text type as a covariate to distinguish estimates based on the original vs. randomized corpora. Finally, we allow text type to interact with the grouping factor. Because we do not have speaker labels for the ADS corpora, only session id was treated as a random intercept. All other aspects of the model including the additional control variables are identical to those reported above.

Results of the models are summarized in Fig. 8 (fuzzy matching) and Fig. 9 (strict matching). Each panel represents the outcome of a single model. Predicted means and confidence intervals are given as points and bars, respectively. Blue corresponds to estimates based on original text and red to estimates based on random text. Numbers on the x-axis correspond to developmental groups, running from 1 (earliest stages of development) to 4 (latest stages of development), with ADS having its own category "adult." Stars indicate the significance of the interaction between text type (random or original) and the developmental grouping factor.

The trends displayed across groups, both developmentally and with respect to the random baseline, mirror closely what we found in the more fine-grained analysis. Chintang variation set proportions tend to increase over development, while English variation set proportions tend to decrease. Thus, the categorical splits that we impose on the continuous data do not seem to obscure the expected trends. English estimates from original text are universally lower than the associated random baselines. Chintang, however, only shows a systematic difference in the fuzzy matching condition. This finding is most likely due to the relatively more complex verbal morphology of Chintang.

The most consistent pattern to emerge is for English. English ADS is uniformly more repetitive than the next closest developmental group. This finding is unexpected, but may be explained by the longer average length of utterances in the BNC. While we control for length of utterance in the regression model, the distribution is strongly bimodal, with CSS centered on a lower average length than ADS. More interesting is the fact that ADS shows a consistently larger difference between the random and original text (statistically significant in all conditions). Recall that this difference reflects the amount of information carried by the average utterance relative to the lexicon and its associated probabilities of occurrence. As the gap widens, more information is being encoded, where information should be understood as the lack of redundancy. The fact that the gap is larger for English ADS means that it is much less repetitive than it could be compared to CSS.[14] In other words, CSS is less informative (i.e., more redundant) given its available repertoire of words. This finding serves as complementary support for the notion that variation sets, and structured repetition more generally, are exaggerated in CSS relative to ADS.

Chintang ADS follows a similar pattern to English in the fuzzy matching condition. The gap between random baseline and original text is larger for adults than for the immediately adjacent developmental groups. However, the ADS sample behaves very similarly to the developmental group 1 (particularly for the PCA-derived index). Nevertheless, the contrast between true and random text is significantly larger in the Chintang ADS sample than in group 1 for all developmental indices in all conditions. Thus, statistically at least, ADS is uniformly less repetitive (more informative) than CSS relative to their respective random potentials.

Finally, we find a general decrease in redundancy between the CSS and ADS bins for both languages, even though they differ in the developmental trends for rates of pure repetition (Chintang increasing and English decreasing). Thus, while repetition and redundancy are both signatures of variation sets, they either operate semi-independently or at

[14] This interpretation was confirmed by examining the normalized ranked frequencies of words in Chintang and English ADS vs. CSS. For both languages, CSS shows a fatter tail for higher-frequency items compared to ADS. More high frequency word types means greater diversity when sampling, hence lower random estimates of repetitiveness (you are less likely to keep drawing the same word). This distribution translates into smaller gaps between the estimates from original and random text.

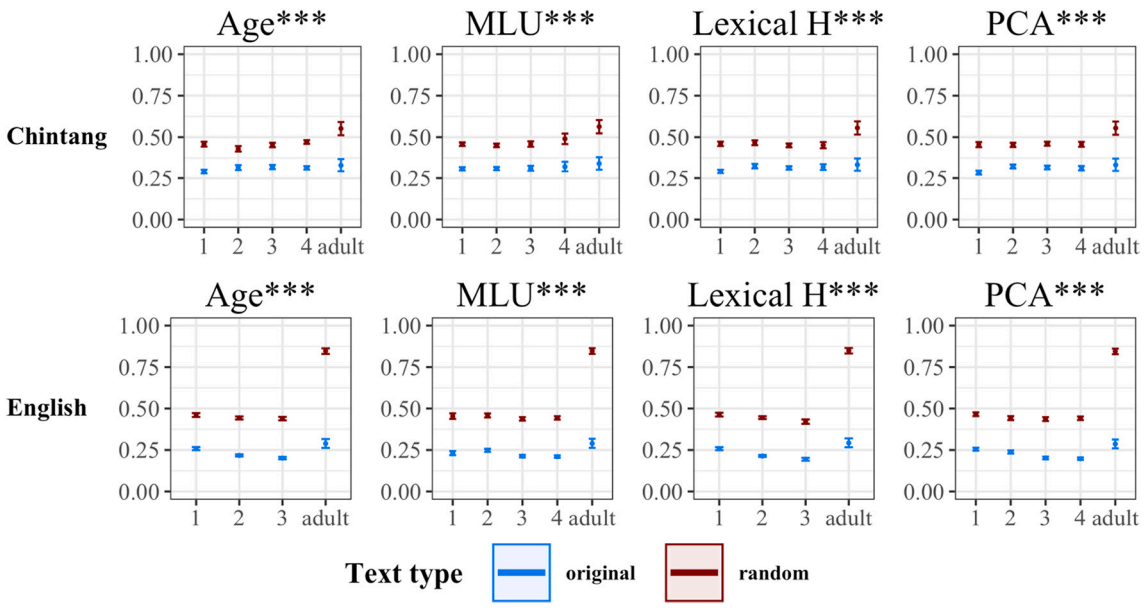


Fig. 8. ADS vs. CSS, words, fuzzy matching.

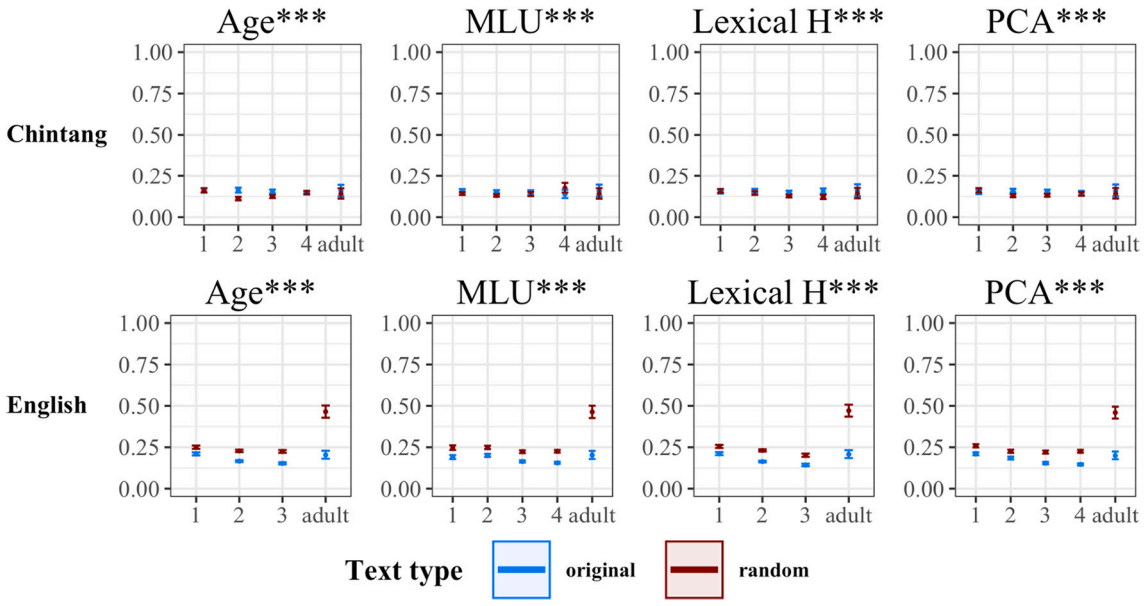


Fig. 9. ADS vs. CSS, words, strict matching.

different time scales. For example, the decrease in redundancy could begin at later stages for Chintang than for English, preceded by early-stage increases (suggested by the diminishing difference between random and true estimates for Chintang CSS). Whatever the explanation, the cross-linguistic decrease in the redundancy of ADS is in line with prior research on variation sets. But now it is apparent that the trend may depend on how one measures repetition, and at what time scale per language.

5. Discussion

Prior research has shown that CSS from many different languages contains repetitive chains of utterances known as variation sets. Recently, the relative pervasiveness of such chains has been estimated automatically by testing the degree of similarity between neighboring utterances in child-surrounding speech. The present study builds on this work in several ways. First, we introduce an automated algorithm that labels utterances as belonging to variation sets according to the following set of parameters: window size and type of matching (strict vs. fuzzy string matching). This algorithm constitutes a more flexible composite of those presented in prior studies. To guard against low-level methodological biases, we systematically varied these parameters and controlled for them in the statistical models presented here. Second, the samples of languages included in previous studies have lacked control for structural/typological, genealogical, and areal biases. We address this point by selecting eight typologically maximally diverse languages from different parts of the world. Third, prior work has relied on larger-scale levels of analysis for matching, such as utterance or word levels, which may not be appropriate for morphologically complex languages. We therefore included a morpheme-level analysis to ensure that we were able to capture repetition of roots in addition to words. Fourth, longitudinal changes in variation sets have most often been observed relative

to the age of the child. However, age is not always a reliable indicator of linguistic development, particularly when comparing multiple children. We therefore compared the behavior of our algorithm across three distinct developmental indices, as well as a measure designed to capture any information shared between those indices. Fifth, no study to our knowledge has established an empirical baseline against which to measure the relative magnitude of the proportion of variation sets given the general statistical properties of a language. We created random versions of the original texts and compared the amount of repetition in each. Sixth and finally, no study to our knowledge has addressed how variation sets in CSS stack up to ADS, nor how this relationship plays out crosslinguistically. We therefore contrasted CSS and ADS for two languages which differ widely, both in terms of several typological parameters (see Appendix) and with respect to the longitudinal behavior of variation sets (Chintang and English). We discuss the consequences of each of these extensions in turn.

Our composite algorithm detected variation sets in all of the languages in our sample. Moreover, we found that each of the parameters had relatively stable effects across languages. Larger windows and fuzzy matching produce larger proportions of variation sets. The results are largely consistent with those reported in prior studies. However, by including a more diverse array of languages, we uncovered some novel trends. Prior work has repeatedly found that languages either tend to decrease or stay the same over time with respect to the observed proportion of variation sets. However, in the present sample two languages, Chintang and Turkish, revealed an unexpected developmental pattern\. For these languages, the prevalence of variation sets increases over time, both at the level of words and morphemes, and for both strict and fuzzy matching. This finding is even more surprising given that these two languages are among the most morphologically complex in our sample. Lexical roots are expected to be repeated if indeed the speech is repetitive. However, the morphological frame surrounding those roots can shift in a number of ways, some of which are obligatory given the interactional context (as in (2) above). There can even be morphophonological processes which alter morphs or obscure morphological boundaries. Yet the positive trend persists, even if we require strict matching over entire words (though somewhat less consistently for Chintang than for Turkish).

These increasing trends clearly challenge the traditional explanation for the longitudinal trends of variation sets, which states that the amount of repetition is inversely related to the socio-cognitive aptitude of the child. Why then should these rates increase for Chintang and Turkish? There are several possibilities.

The simplest and most practically oriented explanation is that these corpora cover the lowest age ranges of any in our sample. Perhaps speech patterns to the youngest children differ from those of interactionally more engaged two- or three-year-olds. This answer could apply especially well to Chintang. In the Chintang community, when children are young, they are not taken seriously as conversational partners. As a result, they hear more speech directed to others than to themselves, at least until they demonstrate proficiency as speakers. Repetition in Chintang CSS might increase over time as caretakers begin to address the child more frequently, thus creating more occasions in which communicative success must be monitored and adjusted for. Perhaps, then, the diminishing repetition that has been observed for other languages is simply pushed further into development for children learning Chintang (an inverted-U trajectory). Interestingly, Grigonytė and Björkenstam (2016) report increasing rates of repetition between their first (0;6–0;11) and second (1;0–1;3) age bins for 3 out of the 10 languages they sampled for which such age ranges. However, unlike the trends demonstrated in the present paper, the three that they present destabilize after the initial increase. This difference may be due to the finer-grained longitudinal analysis pursued here.

Another possibility is that the positive trend truly relates to a special feature CSS in these languages. Perhaps repeating complex forms serves an increasingly important role as children discover more and more of the morphology of their language (especially when the language is morphologically very complex). That is, the adults may implicitly expect more finely articulated comprehension of complex forms, and so offer repetitions and reframings when perceived misunderstandings arise. Consider, for example, that in Chintang, verbal inflections for a single stem result in over 4800 different verb forms because of the language's large number of affixes, and freedom in prefix ordering and verbal compounding in its grammar (Stoll, Mažara, & Bickel, 2017). While "good enough" comprehension (understanding the gist of a message) may suffice early on (as suggested by Frank et al., 2013), increasing demands on accurate interpretation of utterances could yield increased prevalence of variation sets. If this is true, then we have a straightforward prediction. The types of words that participate in variation sets should change as the child ages, with an increasing proportion of words with more complex morphology. Our present methods do not allow us to target the individual forms that lead to matches (and some comparisons between utterances yield many matches). We therefore leave it to future research to test this hypothesis.

Our findings for Turkish are surprising for a different reason, namely, they are at odds with prior research. For example, Grigonytė and Björkenstam (2016) report decreasing proportions in their sample of Turkish CSS. For now, we can only speculate about the source of this difference. First, the corpora in their study cover an older age range (~2;0–4;4) than the corpus analyzed here (~0;7–3;0). As already mentioned with respect to Chintang, the trends we observe could capture a special stage in the development of parent-child interaction, one that is specific to morphologically complex languages. Second, Grigonytė & Björkenstam, 2016 use large bins when computing their average proportions per age group (covering 6–9 months a piece), which could obscure developmental trends that occur within those bins. A related issue is that Grigonytė & Björkenstam do not account for individual differences in the behavior of the specific parent-child dyads or recording sessions. This additional lumping of data could also muddle or distort developmental trajectories. Our models explicitly account for such variability. Third, the corpus analyzed here is not only much larger in terms of the number of utterances analyzed, but also covers four times the number of children. It could be that the decrease observed by Grigonytė & Björkenstam arose from some idiosyncratic combination of features endemic to that sample. Fourth, and finally, at least one of the corpora used by Grigonytė & Björkenstam (Aksu; Slobin, 1982) contains both naturally occurring and experimentally targeted language (i.e., standardized comprehension questions). Those data are therefore not entirely comparable with the purely naturalistic-conversational data analyzed here.

We also tested whether the proportion of variation sets depends on where we look – full words or morphemes. Comparing morphemes uniformly increased the estimated proportion of variation sets. This effect was much stronger in the fuzzy as opposed to the strict matching condition. In some cases, using morphemes instead of words also increased the strength of longitudinal trends. This was particularly true of Turkish and Yucatec. Importantly, the sharpening of the longitudinal trends in these two languages moved in opposite directions (increasing positive correlation for Turkish, negative for Yucatec) and held even with strict matching on noun/verb roots. The morpheme-level analyses thus reveal language-specific layers of repetition that are unavailable at the level of words, and which can be uniquely associated with strict repetition of lexical roots (as opposed to, e.g., fuzzy matching over inflectional morphology in whole words). For other languages, moving to morphemes reduced or even eliminated effects observed at the word level. This change was most pronounced in Sesotho. Perhaps entire words are repeated less and less often, while roots are repeated at steady rates, but in increasingly variable morphological contexts. The word-level analysis should then indicate change over development as a function of increasing morphological flexibility in the use of different roots, but that question lies far outside of the present study. Nevertheless, one thing is clear: the level of analysis at which one measures

repetition is crucial and must be considered in light of the properties of individual languages.

Considering only the analysis of the true texts, languages also differed in how the choice of developmental index (age, MLU, diversity of vocabulary, or the information shared among them) related to changes in the proportion of variation sets. For example, in Chintang, age is the most reliable index for detecting change over development in each condition. MLU, on the other hand, shows no pattern at all. For Japanese, only diversity of vocabulary failed to produce any longitudinal trends. More encouragingly, several of the languages showed convergent effects across all developmental indices. These include Turkish, Russian, and English. The fact that trends are discovered relative to each index, and that these trends are consistent in direction within each language, strongly supports the notion that degree of repetition in adult-child interactions depends on the linguistic ability of the child. But for many other languages, the effectiveness of any single index for detecting longitudinal trends varied both by language and condition. This point again underscores the necessity of applying a wide-ranging approach to the cross-linguistic analysis of variation sets. You have to know where to look, how to measure similarity, and how to measure development for each language independently.

We further introduced a method for generating random texts which share the same underlying statistical properties of the true text. These random baselines were designed to test whether variation sets are surprising at all, or simply a reflex of word or morpheme probabilities. Overwhelmingly, these baselines fell above the true estimates, meaning that for most languages in most conditions, speech was less repetitive than it could have been. We attribute this to the informativity of actual conversation. By not simply reproducing word probabilities in each utterance (i.e., by systematically choosing words that are individually improbable), we increase the amount of information carried by the message; that is, we say something meaningful. On the other hand, repetition can be greater than the random baseline, in which case less probable words are selected repeatedly in sequence for true text relative to what would have been expected by chance. The most likely cause for this effect is a more even distrribution of tokens across types in the higher frequency registers. If there are more high frequency types to choose from, randomly constructed utterances will naturally be less similar (all else being equal). Whatever the cause, these analyses have revealed a new measure of repetition. The absolute value of the difference between the random and true estimate is a measure of redundancy. Positive differences between true and random estimates indicate superabundance of repetition. This situation represents the strongest evidence that CSS is overly repetitive. Negative differences between true and random estimates indicate the extent to which a text is hypo-redundant. Hypo-redundant texts that are closer to the random baseline are more redundant, i.e., "repetitive" in the sense described above. The lower the difference, the more informative the speech. These estimates can evolve over development, including the random baselines themselves. This perspective therefore adds a new perspective on the trustworthiness of developmental trajectories uncovered by automatic variation set extraction algorithms. If there is no difference in the trends, then the ostensible growth or reduction of repetition is most likely due to shifts in the frequency distribution of tokens in the text.

We indeed found a few cases in which an otherwise reliable longitudinal trend was in fact indistinguishable from natural changes in repetitiveness due to changes in the frequency profile of words/morphemes. Inuktitut appeared to show a positive longitudinal trend for the PCA-derived index, but this trend mirrored almost exactly the increase observed for the randomized text. Hence, we should not attribute these changes to anything other than basic lexical distributions. One use for the random baseline is thus to weed out spurious correlations between repetition and development.

We also found evidence of the converse situation: there were cases in which the longitudinal trends for true and randomized text differed, but no effect was observed for the true text alone. For example, with strict matching of morphemes, Yucatec showed no effect of MLU on its own, but did show an interaction between MLU and text type. Random text increased over time in variation set proportions, while true text did not. The result was an increase in the distance between the two, with the random baseline reaching values increasingly above those of the standard. By the logic presented above, even though the actual repetition has not increased in response to increasing MLU, the amount of redundancy has decreased. This insight is crucial: it means that sequential repetition alone is not the only meaningful dimension of variation sets. We must also consider redundancy in terms of violation of expectation. Specifically, even if the rates of repetition are constant across samples, this constancy may be shaped by short-scale recurrence of lower-frequency items. This explanation implies that for such corpora, we should observe increasingly shorter chains of variation sets driven by increasingly lower-frequency items. Our current approach does not allow us to measure chain length, though we note that these effects hold when window size (i.e., maximal chain length according to our algorithm) has been held constant. We leave it to future research to determine the root causes of these various types of null effect (null simple effect, but significant with interaction; null interaction, but significant with simple effect).

In our final analysis, we tested whether CSS is especially repetitive compared to ADS. Should we count variation sets among the distinctive features of adult-child interaction, at least in terms of their relative prevalence, or not? We compared two typologically distinct languages: English and Chintang. Our analysis revealed two important things. First, the longitudinal trends for Chintang gradually approximate the adult standard for all indices, but this is not true for English. For the latter, rates of repetition in CSS gradually decrease, while those for ADS rest substantially higher than even the earliest stages. Moreover, the difference in repetitiveness between English ADS and CSS is much greater than it is for Chintang, regardless of the developmental stage of the child. This could be due to a fundamental difference in average utterance length between the corpora (Chintang ADS: mean words = 2.76, SD = 1.97; English ADS: mean words = 9.41, SD = 13.41). A qualitative examination of the BNC Spoken corpus revealed that the unit of chunking was longer than that used for Chintang, something closer to a turn than an individual utterance. For example, a single "utterance" could contain multiple sentences. However, Chintang is expected to have shorter utterances by word than English simply because English is morphologically more analytic. While we cannot definitively answer the question here, we emphasize the basic point that automatic detection of repetition is sensitive to both language-internal features, such as morphological complexity, and/or corpus-specific features, such as how utterances are chunked into transcription units. Second, both Chintang and English showed bigger differences between true and random estimates for ADS than CSS. ADS is therefore more informative on average than CSS in these languages, despite the differences outlined above. This trend is consistent between languages, and in line with the expected decreases in repetition based on prior work on the assumption that less repetitive speech is more informative. Thus, repetition behaves in unique ways for CSS and ADS, both in the levels of pure repetition and the degree to which these levels of repetition are informative. Moreover, cross-linguistic differences suggest that the effects of repetition and redundancy may play out on different time scales for different languages. To test this possibility, we need corpora covering later stages of development (e.g., between the ages of 5 and 18+).

While these results are compelling, our automated approach is not without its shortcomings. For example, we necessarily underestimate the true amount of repetition. We currently only measure whether an utterance belongs to *any* variation set. But any utterance may participate in several distinct variation sets within the same window. Consider the following simple sequence:

1. A B C
2. A D E

3. C D F

Our proportional measure would say that 100% of the utterances belong to variation sets. But in actuality, each of the utterances belong to two variation sets (1 with 2 and 3; 2 with 1 and 3; 3 with 1 and 2), each of which centers on one of three elements (A, C, or D). The proportion of utterances as implemented here cannot capture this interweaving of repetition. Additional measures are therefore necessary. For example, one could calculate the average number of variation sets per utterance (in the above example, 2). This measure would capture the density of variation sets as opposed to their prevalence and therefore give a better impression of the network of repetitions encountered in CSS. Another option would be to identify variation sets through cluster analysis and then to count the number of clusters found in each session.

Another source of underestimation comes from the fact that our algorithm cannot capture phonologically obscure relationships between words or morphological roots. For example, repetition of radically suppletive forms, such as English *go* and *went*, would not register as a match. One way of handling this issue would be to apply the algorithm over lemmas (essentially words stripped of inflection; *go* and *went* share a lemma GO). However, there are two issues. First, lemmatizing the tokens presupposes that the child can connect these disparate surface forms already, and so pushes the question one step back. Second, lemmatization itself is not straightforward for many languages. Consider Chintang, in which verbs can carry multiple, morphologically discontinuous roots. Which one(s) should be built into the lemma? It is therefore unclear that any automated extraction algorithm could account for these relationships. Another way to approach this problem would be to examine just those sequences in which target suppletive forms are repeated. It is possible that these forms are linked by variation sets of a different kind, sets where the targets vary and the context remains stable. In this case, local chains of varying suppletive forms would be attended by higher degrees of repetition in the embedding contexts. If most of the sentence is repeated, but the target forms are allowed to vary, the child may have an easier time connecting them. Similar mechanisms have been proposed for the bootstrapping of lexical categories under the label "frequent frames" (Mintz, 2003; Moran et al., 2018). We leave it to future research to explore specifically whether phonologically divergent variants of the same underlying form can be learned through variation sets, or through some combination of variation sets and other statistical properties of text.

The data themselves also limit the precision with which we can estimate the proportion of variation sets. What we refer to as utterances actually fall somewhere between a full turn at talking and a coherent syntactic unit. However, speech provides other cues to segmentation, namely, prosodic cues. Speech comes packaged in prosodically well-defined bursts known as intonation units (Du Bois, Schuetze-Coburn, Cumming, & Paolino, 1993; Chafe, 1994). Our current utterance units are typically longer than a single intonation unit. Hence, our units of comparison may be too large. As a result, we may omit relevant matches because they fall within our current utterance units, but would otherwise be split among two or more intonation units. However, segmenting the corpora into more total units also increases the denominator of our critical proportion. It is therefore unclear whether the estimated proportions would rise or fall, and whether these relationships would depend on other factors, such as the typological properties of the language or whether we consider morphemes or words.

Beyond the automatic variation set extraction algorithms and how they interact with morphological and lexical complexity, other variables may in general affect the proportions of utterances belonging to variation sets. For example, socially- or culturally-driven differences between speaker communities has been shown to be a significant determinant of variation sets. Tal, Arnon, Bertolini, and Kaplan (2018) find that children learning English or Hebrew from high socio-economic status (SES) backgrounds receive more variation sets compared to children from low SES backgrounds. After controlling for the difference in the number of words, they find that higher SES language samples from English and Hebrew have statistically significantly higher proportions of variation sets than lower SES samples. However, in some of the cultures studied in this paper such differences are irrelevant. In Chintang for instance the population lives from subsistence farming and SES differences are not comparable to WEIRD cultures, as for instance, studied in Tal et al. (2018).

Another factor to consider is how behavioral standards for adult-child interaction shape the prevalence of variation sets across cultures. Cultures differ enormously in how much and how caregivers communicate with children of different ages (Casillas, Brown, Levinson, & C., 2020; Casillas, Brown, Levinson, & C., 2021; Cristia, Dupoux, Gurven, & Stieglitz, 2019; Keller et al., 2006; Lieven & Stoll, 2013; Ochs & Schieffelin, 1995). Moreover, cultures differ in their preferred models of discourse, which may impact the typical prevalence of variation sets. For example, Mayan languages have been documented to involve a high degree of full and partial repetition (Brody, 1986; Brown, 1999). Research into Mayan child-directed speech suggests that there are also developmental trends in this behavior. For example, K'iche'-speaking mothers repeat themselves more when talking to children, but they repeat things that others have said more when speaking to adults (Pye, 1986). The prompting routines described above for Yucatec Maya also give way to higher degrees of partial or full repetition in the speech between adults, suggesting the possibility of a U-shaped development in the prevalence of variation sets. Going forward, computational approaches for identifying variation sets must consider the full scope of social, ethnographic, and discourse factors, as well as the relationships between child-directed and adult-directed speech, where sufficient documentation exists.

6. Conclusion

Variation sets are a ubiquitous feature of child-surrounding speech (CSS). Detecting these variation sets automatically has drawn much attention in recent years. Here we show through a myriad of automatic extraction algorithms and statistical analyses on a typologically maximally diverse sample of languages that variation sets in CSS can either increase or decrease in frequency as a function of the child's age. Furthermore, the present study adds several new dimensions to consider in the analysis and interpretation of variation sets.

First, we must consider carefully how a language is structured before submitting it to automatic analysis. Whereas morphologically analytical languages like English favor the repetition of whole words within varying phrasal or sentential contexts, synthetic languages like Turkish favor repetition of roots within varying morphological contexts. These differences affect our ability to detect repetition, but on a deeper level, they seem to correspond to real differences in how repetition manifests longitudinally. The fact that some morphologically complex languages in this study show increasing rates of repetition was entirely unexpected and deserves further scrutiny. We suspect that this increase is tied into a larger inverted-U developmental trajectory, but the answer will have to wait until we have sufficiently large and longitudinally broad cross-linguistic corpora.

Second, we cannot simply treat age as the standard measure of linguistic aptitude, nor the measure that is most appropriate for any given sample of children, or any language for that matter. We have tested two additional measures, MLU and lexical/morphological diversity, but this hardly exhausts the full gamut of developmental indices. Future research should continue to look for convergence across multiple developmental indices before drawing specific conclusions about how variation sets change in response to the growing linguistic aptitude of the child.

Third, random baselines must be established against which both point estimates and developmental trajectories can be compared. The difference in repetitiveness between random and true text offers a new perspective on variation sets: when true and random estimates are

closer, the speech is more redundant given the underlying frequency-rank distribution of words/morphemes. Redundancy and repetition go hand in hand, as both mark pockets of less informativity. In other words, they are two sides of the same coin. However, in some cases one is easier to detect than the other with our algorithm. The relative usefulness of each thus needs to be evaluated further using different algorithms, developmental indices, and languages. At a more practical level, we have introduced only one approach for generating random text, which involves randomly sampling from an empirically derived frequency distribution per text. But there are many other possible techniques, such as randomizing across multiple levels (characters within words, words within utterances, utterances within texts, and so on). Based on the present results, we suggest that the most desirable approach would be one that guarantees a fixed direction of divergence (positive or negative) in the proportion of variation sets between true and random text. That way, we have a set scale against which to judge the relative effect size.

Finally, our analysis showed that we must compare the behavior of ADS and CSS, and that this comparison must include some form of random baseline. We assumed that CSS would gradually approximate ADS, which was true for Chintang but not English. This difference could be due to several factors, including how speech was transcribed and segmented. But one fact is clear: ADS was reliably less redundant than CSS in both languages. The scale of this difference offers another aggregate view of the degree of repetition in CSS. Notably, this difference matches prior findings, and conflicts with the patterns of pure repetition in CSS. Further work should expand this style of analysis to new languages and new corpora of the languages studied here, especially corpora with a broader coverage of older age ranges. In this way, we can begin to understand the possible effects of typological features and transcriptional conventions.

We propose that by considering these many additional dimensions, the automatic detection of variation sets will benefit both in terms of precision and interpretability. Furthermore, our findings are relevant for theory in two ways. First, a comprehensive theory of variation sets must explain why repetition decreases for some languages but increases for others over time. Second, the property of redundancy must be considered in complement to pure repetition. By hypothesis, if variation sets are prevalent, then less information is transmitted on average per utterance. With these additional tools, we can sharpen our understanding of CSS and how it supports language acquisition across languages and cultures.

Acknowledgements

Nicholas A. Lester, Steven Moran, and Sabine Stoll were supported by the project 'Acquisition processes in maximally diverse languages: Min(ding) the ambient languages (ACQDIV),' which received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7-2007-2013) (Grant agreement No. 615988; PI Sabine Stoll). Steven Moran received funding from the Swiss National Science Foundation (Grant No. PCEFP1_186841). We would also like to thank the attendees of the 43rd Annual Boston University Conference on Language Development, as well as three anonymous reviewers for their insightful comments and suggestions.

Appendix 1: Typological parameters and feature values for the languages in our sample based on Stoll and Bickel (2013)

Language	Verb position	Verbal synthesis	Nominal synthesis	Syncretism	Verb agr.	Possessive agr.	Case A. vs P.	Agr. split ergative	Case split ergative	Polyexponence	Inflectional compactness
Chintang	V = 3	High	2	Some	Some	Some	Some	Low	Low	Some	Distributive
English	V = 2	Low	1	Some	Some	Some	None	Low	Low	Some	Cumulative
Inuktitut	V = 3	Medium	3	Some	Some	Some	Some	Low	High	Some	Distributive
Japanese	V = 3	Low	1	None	None	Some	Some	Low	Low	Some	Cumulative
Russian	V = 2	Low	2	Some	Some	None	Some	Low	Low	Some	Cumulative
Sesotho	V = 2	Low	1	Some	Some	None	None	Low	Low	Some	Distributive
Turkish	V = 3	Medium	3	None	Some	None	Some	Low	Low	Some	Separative
Yucatec	Free	Low	2	None	Some	Some	None	Medium	Low	None	Separative

References

Aguado-Orea, J. (2004). *The acquisition of morpho-syntax in Spanish: Implications for current theories of development*. Unpublished PhD Dissertation. Nottingham, UK: University of Nottingham.

Allen, S. E. M. (1996). *Aspects of argument structure acquisition in Inuktitut*. Amsterdam: John Benjamins.

Allen, S. E. M. (2021). *Allen Inuktitut Child Language Corpus*. Unpublished.

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

Bannard, C., & Lieven, E. (2009). Repetition and reuse in child language learning. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. M. Wheatley (Eds.), *2. Formulaic Language: Acquisition, loss, psychological, reality, and functional explanations* (pp. 299–321). Philadelphia: John Benjamins.

Bard, E. G., & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language, 10*(2), 265–292. https://doi.org/10.1017/S0305000900007777

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., Zúñiga, F., & Lowe, J. B. (2017). The AUTOTYP typological databases. Online https://github.com/autotyp/autotyp-data/.

Bickel, B., Stoll, S., Gaenszle, M., Rai, N. K., Lieven, E., Banjade, G., Bhatta, T. N., et al. (Eds.). (2011). *Audiovisual corpus of the Chintang language, including a longitudinal corpus of language acquisition by six children: Ca. 650,000 words transcribed and translated, of which ca. 450,000 glossed, plus paradigm sets and grammar sketches, ethnographic descriptions, photographs*. Nijmegen, Leipzig: DoBeS, Universität Leipzig. http://www.mpi.nl/DOBES.

Björkenstam, K. N., & Wirén, M. (2014). Multimodal annotation of synchrony in longitudinal parent–child interaction. In *Proceedings of the 10th workshop on multimodal corpora: Combining applied and basic research targets*. Reykjavik, Iceland: ELRA.

Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Branigan, H. P., & Messenger, K. (2016). Consistent and cumulative effects of syntactic experience in children's sentence production: Evidence for error-based implicit learning. *Cognition, 157*, 250–256.

Brodsky, P., Waterfall, H. R., & Edelman, S. (2007). Characterizing motherese: On the computational structure of child-directed language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Brody, J. (1986). Repetition as a rhetorical and conversational device in Tojolabal (Mayan). *International Journal of American Linguistics, 52*(3), 255–274.

Brown, P. (1999). Repetition. In A. Duranti (Ed.), *9. Special issue of the journal of linguistic anthropology*. Oxford: Blackwell Publishers.

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

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*, 121–170.

Casillas, M., Brown, P., Levinson, S., & C.. (2020). Early language experience in a Tseltal Mayan village. *Child Development, 91*(5), 1819–1835.

Casillas, M., Brown, P., Levinson, S., & C.. (2021). Early language experience in a Papuan community. *Journal of Child Language, 48*(4), 792–814.

Chafe, W. (Ed.). (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood: Ablex.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chao, A., Wang, Y. T., & Jost, L. (2013). Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution, 4*, 1091–1100.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child- directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development, 90*(3), 759–773.

Demuth, K. (2015). Demuth Sesotho Corpus. Online http://childes.talkbank.org/access/Other/Sesotho/Demuth.html.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online https://wals.info/.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (1993). Outline of discourse transcription. In J. A. Edwards, & M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research* (pp. 45–89). Hillsdale: Lawrence Erlbaum.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development, 9*(1), 1–24.

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science, 31*(2), 311–341.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*(2), 183–206.

Goodman, J. C., Philip, S. D., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35*, 515–531.

Grigonytė, G., & Björkenstam, K. N. (2016). Language-independent exploration of repetition and variation in longitudinal child-directed speech: A tool and resources. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016 130, 41–50*. Linköping: University Electronic Press.

Haiman, J. (1997). Repetition and identity. *Lingua, 100*(1–4), 57–70.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology, 22*(2), 155.

Hoff-Ginsberg, E. (1990). Maternal speech and the child's development of syntax: A further look. *Journal of Child Language, 17*(1), 85–99.

Hoiting, N., & Slobin, D. I. (2002). What a deaf child needs to see: Advantages of a natural sign language over a sign system. In R. Schulmeister, & H. Reinitzer (Eds.), *Progress in sign language research. In honor of Siegmund Prillwitz / Fortschritte in der Gebärdensprachforschung. Festschrift für Siegmund Prillwitz* (pp. 268–277). Signum.

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology, 2*, 17.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45*, 337–374.

Jakobson, R. (1966). Grammatical parallelism and its Russian facet. *Language, 42*(2), 399–429.

Jancso, A., Moran, S., & Stoll, S. (2020). The ACQDIV corpus database and aggregation pipeline. In *Proceedings of the Twelfth International Conference on Language Re- sources and Evaluation (LREC 2020)*. Online http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.20.pdf.

Jurafsky, D., & Martin, C. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice Hall.

Keller, H., Lamm, B., Abels, M., Yovsi, R., Borke, J., Jensen, H., , … Tomiyama, J., et al. (2006). Cultural models, socialization goals, and parenting ethnotheories: A multicultural analysis. *Journal of Cross-Cultural Psychology, 37*(2), 155–172.

Krajewski, G., Lieven, E. V. M., & Theakston, A. L. (2012). Productivity of a polish child's inflectional noun morphology: A naturalistic study. *Morphology, 22*, 9–34.

Küntay, A. C., Koçbaş, D., & Taşçı, S. S. (2021). *Koç University Longitudinal Language Development Database on language acquisition of 8 children from 8 to 36 months of age*. Unpublished.

Küntay, A. C., & Slobin, D. I. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In D. I. Slobin, J. Guo, & A. Kyratzis (Eds.), *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp* (pp. 265–286). Hillsdale, NJ: Erlbaum.

Küntay, A., & Slobin, D. I. (2002). Putting interaction back into child language: Examples from Turkish. *Psychology of Language and Communication, 6*(1), 5–14.

Lester, N. (2018). *The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition: UC Santa Barbara PhD dissertation*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *In Soviet physics doklady, 10*(8), 707–710.

Lieven, E., & Stoll, S. (2013). Early communicative development in two cultures: A comparison of the communicative environments of children from two cultures. *Human Development, 56*(3), 178–206.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Desiging and building a corpus of everydqy conversations. *International Journal of Corpus Linguistics, 22*(3), 319–344.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91–117.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek, & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 31–63). Oxford: Oxford University Press.

Miyata, S. (2012). Japanese CHILDES: The 2012 CHILDES manual for Japanese. http://www2.aasa.ac.jp/people/smiyata/CHILDESmanual/chapter01.html.

Moran, S., Blasi, D. E., Schikowski, R., Küntay, A. C., Pfeiler, B., Allen, S., & Stoll, S. (2018). A universal cue for grammatical categories in the input to children: Frequent frames. *Cognition, 175*, 131–140.

Moran, S., Schikowski, R., Jung, D., & Stoll, S. (2021). *ACQDIV Corpus database user manual*. Online. https://github.com/acqdiv/corpus_manual/blob/master/corpus_manual.pdf.

Moran, S., Schikowski, R., Pajović, D., Hysi, C., & Stoll, S. (2016). The ACQDIV database: Min(d)ing the ambient language. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/pdf/1198_Paper.pdf.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, H. (2004). Putting the bits together: An information-theoretic perspective on morphological processing. *Cognition, 94*(1), 1–18.

Naigles, L., & Hoff-Ginsburg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language, 25*, 95–120.

Ochs, E., & Schieffelin, B. B. (1995). The impact of language socialization on grammatical development. In P. Fletcher, & B. MacWhinney (Eds.), *The handbook of child language* (pp. 73–94). Oxford: Basil Blackwell.

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition, 109*(3), 423–430.

Pfeiler, B. (2021). *Pfeiler Yucatec Child Language Corpus*. Unpublished.

Pye, C. (1986). Quiché Mayan speech to children. *Journal of Child Language, 13*(1), 85–100.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Ratcliff, J. W., & Metzener, D. E. (1988). Pattern-matching-the gestalt approach. *Dr Dobbs Journal, 13*(7), 46.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425–469.

Rowland, C. F., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language, 33*(04), 859–877.

Rowland, C. F., Pine, J. M., Lieven, E., & Theakston, A. L. (2005). The incidence of error in young children's wh-questions. *Journal of Speech, Language, and Hearing Research, 48*(2), 384–404.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition, 69*(2), 105–134.

Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2006). Structural priming as implicit learning in language acquisition: The persistence of lexical and structural priming in 4-year-olds. *Language learning and development, 2*(1), 27–49.

Schwab, J. F., & Lew-Williams, C. (2016). Repetition across successive sentences facilitates young children's word learning. *Developmental Psychology, 52*(6), 879–886. https://doi.org/10.1037/dev0000125

Schwab, J. F., & Lew-Williams, C. (2017). Discourse continuity promotes children's learning of new object labels. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 3101–3106). Austin: Cognitive Science Society.

Slobin, D. (1982). Universal and particular in the acquisition of language. In E. Wanner, & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 128–172). New York: Cambridge University Press.

Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German and English child-directed speech. *Cognitive Science, 33*, 75–103.

Stoll, S., & Bickel, B. (2013). Capturing diversity in language acquisition research. In B. Bickel, L. A. Grenoble, D. A. Peterson, & A. Timberlake (Eds.), *Language typology and historical contingency: Studies in honor of Johanna Nichols* (pp. 195–260). Amsterdam: Benjamins.

Stoll, S., Bickel, B., Lieven, E., Banjade, G., Bhatta, T. N., Gaenszle, M., … Rai, N. K. (2012). Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language, 39*(2), 284–321.

Stoll, S., Mažara, J., & Bickel, B. (2017). The acquisition of polysynthetic verb forms in Chintang. In M. Fortescue, M. Mithun, & N. Evans (Eds.), *The Oxford handbook of polysynthesis. Oxford Handbooks Online* (pp. 495–514). Oxford University Press.

Stoll, S., & Gries, S. Th. (2009). How to measure development in corpora? An association strength approach. *Journal of Child Language, 36*, 1075–1090.

Stoll, S., & Meyer, R. (2008). *Audio-visional longitudinal corpus on the acquisition of Russian by 5 children*.

Tal, S., Arnon, I., Bertolini, A. B., & Kaplan, M. J. (2018). SES differences in the communicative functions of variation sets. In *BUCLD 42: Proceedings of the 42nd annual Boston University conference on language development* (pp. 736–749).

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127–152.

Vasilyeva, M., & Waterfall, H. (2012). Beyond syntactic priming: Evidence for activation of alternative syntactic structures. *Journal of Child Language, 39*(02), 258–283.

Vogt, P., & Lieven, E. (2010). Verifying theories of language acquisition using computer models of language acquisition. *Adaptive Behavior, 18*(1), 21–35.
Waterfall, H. (2006). *A little change is a good thing: Feature theory, language acquisition and variation sets*. University of Chicago. PhD dissertation.
Wirén, M., Kristina, N. K., Björkenstam, G. G., & Cortes, E. E. (2016). Longitudinal studies of variation sets in child-directed speech. In *The 54th annual meeting of the association for computational linguistics, Berlin, Germany, August 11, 2016* (pp. 44–52). Association for Computational Linguistics.