Variation Sets in Maximally Diverse Languages

Steven Moran, Nicholas A. Lester, Heath Gordon, Aylin Küntay, Barbara Pfeiler, Shanley Allen, and Sabine Stoll

1. Introduction

Child-directed speech (CDS) has been shown to facilitate language learning through different structural features and frequently occurring patterns, including statistical regularities of isolated words (Lew-Williams, Pelucchi, & Saffran, 2011), adjacent dependencies (Redington, Chater, & Finch, 1998), non-adjacent dependencies (Mintz, 2003; Moran et al., 2018), and sentence frames (Fernald & Hurtado, 2006). Additionally, CDS is highly repetitive with frequent repetitions of, for instance, utterance initial constructions (Cameron-Faulkner, Lieven, & Tomasello, 2003; Stoll, Abbot-Smith, & Lieven, 2009). An important question for language acquisition studies is not only whether repetitiveness in structures appears in CDS across languages but also how these structures are distributed across contexts. Talking to young children introduces both the challenge of capturing their attention and getting a message across to a not-yet competent speaker. This is presumably one of the main reasons for ubiquitous repetitions and reformulations of lexical items and constructions.

Here we focus on what have been referred to as variation sets (Hoiting & Slobin, 2002; Küntay & Slobin, 1996, 2002; Onnis, Waterfall, & Edelman, 2008; Waterfall, 2006). Variation sets are typically defined as repetitions of individual lexemes in close proximity occurring in interactional units with a constant communicative intent. In previous studies, the definition of proximity, and the types of lexical items studied, varied across languages. However, the findings have been largely convergent: proportions of variation sets in CDS to young children are relatively large, and they tend to decrease over time. Moreover, variation sets support earlier word production in child speech (Waterfall, 2006),

* Steven Moran, University of Zurich, steven.moran@uzh.ch. Nicholas A. Lester, University of Zurich. Heath Gordon, University of Zurich. Aylin Küntay, Koç University, Barbara Pfeiler, Universidad Nacional Autónoma de México, Shanley Allen, University of Kaiserslautern, Sabine Stoll, University of Zurich. The research leading to these results received funding from the European Union's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 615988 (PI Sabine Stoll). Many thanks to Elena Lieven and several anonymous BUCLD reviewers. Author contributions: SM, SST designed the research. SM, HG performed the research. SM, NAL, HG performed statistical analysis. SST, AK, SA, BP provided data. SM, NAL, HG, SST wrote the paper.

 Proceedings of the 43rd Boston University Conference on Language Development, ed. Megan M. Brown and Brady Dailey, 427-440. Somerville, MA: Cascadilla Press.

though perhaps not morphological productivity (e.g., for action nouns in Yucatec; Pfeiler, 2017).

Küntay and Slobin (1996, p. 267) demonstrated the notion of variation set with English data from a father prompting the memory of his child aged 2;3. The variation set revolves around the verb 'to see' and the question 'Who did we see?'. To get his message across the father repeats and rephrases the same question several times:

(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**?

The repetition of lexical items and their embeddings in different constructions strongly depends on grammatical features of a language. Küntay and Slobin (1996) compared variation sets in English and Turkish child-directed speech. Turkish has complex noun and verb morphology, high rates of nominal ellipsis, and a freer word order than English. The database they analyze consists of 3,167 utterances from a mother to her child aged 1;8-2;3. An example variation set from Turkish is given in (2) below (Küntay & Slobin, 1996, p. 270):

(2)

- **Ver** el-ler-in-i.
  give hand-PL-POSS.2SG-ACC
  'Give (me) your hands.'
- El-ler-in-i **ver**-ir-mi-sin.
  hand-PL-POSS.2SG-ACC **give**-AOR-Q-2SG
  'Will you give (me) your hands?'
- El-ler-in-i **ver**.
  hand-PL-POSS.2SG-ACC **give**
  'Give (me) your hands.'

Here, the morpheme *ver,* the stem of the verb 'give' appears in multiple contexts, with morphological, lexical, and syntactic variation. Note also that this variation set intersects with others, including those centered on *el* 'hand'*, -ler* PL*,* and the complex expression *eller* 'hands.' From the side of the speaker, variation sets are useful to get a message across. But they also seem to have an advantage for the language learner. Küntay & Slobin (1996) suggest that the entire set of grammatical cues provided by these variable environments are necessary for Turkish children to track lexical items across varying utterance positions, with different associated agglutinated morphemes. For example, the first and third instances of *ver*, which are bare stems and occur without additional morphemes, while the second instance surfaces as a complex verb form with three inflectional

morphemes. Onnis et al. (2008, p. 424) note that the operative characteristic of a variation set is that it entails a local mechanism of alignment, which allows even memory-limited learners to compare and discover structure in their input. This observation is supported by other work, in which repetitiveness in the input to children has been shown to help in the segmentation of speech (Bard & Anderson, 1983) and to predict syntactic development (Hoff-Ginsberg, 1986, 1990; Waterfall, 2006). Also, the temporal proximity and continuity of repetitions in language have been reported to create supportive contexts for partial utterance understanding leading to a more complete understanding of the respective constructions (Frank, Tenenbaum, & Fernald, 2013).

Küntay & Slobin (1996) note that the comparison of variation sets across languages can be challenging especially if morphologically complex languages are involved. As such, variation sets have been defined differently by different researchers, specifically in light of operationalizing algorithmic approaches to automatically extract variation sets from corpora. Thus, the goals of our study are to (i) operationalize variation sets and present several methods to automatically extract them from corpora, (ii) test whether variation sets can be found in typologically radically different languages, and (iii) test whether they vary as a function of age, i.e. whether they decrease in the input over time, as suggested by Waterfall et al. (2010) and Wirén et al. (2016).

2. Background

Various definitions and methods to automatically extract variation sets from child language acquisition corpora have been proposed. 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) collected by Goldin-Meadow, Huttenlocher, & Levine (2002-2007).[1] She finds a decrease in the proportion of utterances in variation sets as a function of age, i.e. a decrease from 17% to 12% from 1;2 to 2.6 years.

Onnis et al. (2008) report a 27.9% overall rate of utterances in variation sets in the Lara corpus (Rowland & Fletcher, 2006; Rowland, Pine, Lieven, & Theakston, 2005) using the definition of variation sets given in Waterfall (2006). They also loosened the criteria for defining variation sets, including any single-word overlap (not just nouns and verbs). This approach yielded 58.6% of the utterances as part of variation sets. Further, they found that 34.9% of unique words surfaced 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,

1. National Institutes of Health grant # POl HD40605.

pronouns, auxiliaries, WH-questions, proper names and a set of function words. With this definition, Brodsky et al. (2007) reanalyzed 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) and find that 18.3% of the words occur in variation sets. The divergent results obtained illustrate that differing definitions of variation sets and their operationalizations have an impact on how many variation sets are identified.

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 did stepwise comparison 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. They then 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 CDS. The proportion of verbatim repetitions also decreased.

Grigonytė & 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 below): the Ratcliff–Obershelp pattern recognition method and the Python module *difflib*[2], which is a library that provides string similarity measures, including edit distance (Levenshtein, 1966). They compare their results using these approaches against the Swedish gold standard database (Wirén et al., 2016). Then they apply them to 26 corpora[3] in the CHILDES database (MacWhinney, 2000). Their open-source Varseta tool[4] achieves (semi-successful) precision and recall figures when applied to gold standard datasets of Swedish and French (Grigonytė & Björkenstam, 2016). F-scores perform relatively poorly across the board, regardless of the

2. https://docs.python.org/2/library/difflib.html

3. Afrikaans, Cantonese, Catalan, Chinese, Croatian, Danish, Dutch, English, Estonian, Farsi, French, German, Greek, Hebrew, Hungarian, Indonesian, Irish, Italian, Japanese, Portuguese, Russian, Spanish, Tamil, Thai, Turkish, Welsh.

4. https://github.com/ginta-re/Varseta

anchor or incremental stepwise analysis. Interestingly, they do not find a decrease in the proportion of exact repetitions for most of the languages in their sample. However, there are exceptions, including Swedish, Danish, English, Russian, Cantonese, Japanese, Thai, Welsh, Estonian, Hebrew, and Tamil. A general trend they observe in their data, given their approaches for variation set extraction, is that the proportion of exact repetitions observed is less than the proportion of utterances they find in variation sets. Lastly, Grigonyte & Bjorkenstam (2016) observe that the proportion of utterances belonging to variation sets decreases as the target child gets older in all languages, except for for French and Portuguese. Also, for nearly half of the languages in their sample, there is a decrease in the proportion of exact repetitions. However, this trend is not supported in Swedish, Danish, English, Romanian, Cantonese, Japanese, Thai, Welsh, Estonian, Hebrew, and Tamil.

In the next section, we introduce a comprehensive approach to automatic, cross-linguistic extraction of variation sets in child-directed speech. We build on prior studies by (i) comparing the effects of several parameters on the observed proportion of variation sets across these languages and (ii) including a more diverse array of languages.

3. Methods

Here we combine features of previous analyses to produce a general procedure for automatically extracting variation sets from text-based corpora, which takes into account the parameters used in previous studies. We then evaluate and compare the performance of several variation sets operationalizations across different languages and age ranges, which we report in the result section. The parameters include

- Window utterance size
- Type of pairwise comparison
    - Anchor
    - Incremental
- Number of matches
- Type of match (exact or fuzzy)
- Comparison type
    - Nouns and verbs
    - Words and morphemes

The **window utterance size** is the number of consecutive utterances that we use to make our pairwise comparisons. Utterances are compared pairwise, but the method of finding utterances to compare can either be by using an anchor or incremental method. The **anchor** strategy measures pairwise similarity of utterances in relation to an anchor utterance given some number of utterances per so-called window, e.g. with a utterance window of 5: 1-2, 1-3, 1-4, 1-5. In practice this means comparing the first utterance with the successive utterances in the

window. The **incremental** strategy measures pairwise similarity of utterances in successive utterances given an utterance window, e.g. 1: 1-2, 2-3, 3-4.

The number of matches specifies the amount of tokens that must overlap between the utterances in order to be considered a variation set. Exact matches of words are considered, as are fuzzy matches, where a threshold for edit distance allows us to count a partial match (e.g. 'dog' and 'dogs'). The **comparison type** dictates at what morphological level the analysis is performed. For example, we may consider matches between complete word forms (identity match, e.g., 'cat' with 'cat'). Or we may further condition these matches to consider only specific parts speech, e.g., **nouns** or **verbs**, so that 'bank' (noun; financial institution) and 'bank' (verb; participate in a financial transaction at a bank) would be considered separately. Alternatively, we can match on sub-lexical items, i.e., **morphemes**.

4. Data

As input to our methods, we use the ACQDIV database (Moran, Schikowski, Pajović, Hysi, & Stoll, 2016; Stoll & Bickel, 2013), which consists of nine longitudinal child language acquisition corpora from typologically maximally diverse languages: Chintang (Stoll et al., unpublished), Cree (Brittain, 2015), Indonesian (Gil & Tadmor, 2007), Inuktitut (Allen, unpublished), Japanese (Miyata, 2012), Russian (Stoll & Meyer, 2008), Sesotho (Demuth, 2015), Turkish (Küntay, Koçbaş, & Taşçı, 2015), and Yucatec (Pfeiler, unpublished). These corpora include transcribed speech produced by 46 target children between ages one and six and from surrounding adult and other children. For comparability with previous studies, we also added the English Manchester corpus to our study (Theakston, Lieven, Pine, & Rowland, 2001). Here, we consider only the speech produced by adults in the presence of children, of child-surrounding speech (CSS). CSS includes both child-directed speech and utterances that may have been overheard by the children. Table 1 summarizes the data sample that we used. Session counts reflect the number of independent recordings made across all children. Utterance counts are based on segmentations of the text made by expert transcribers into coherent turns at speech. They correspond roughly to clauses but do not strictly align with any syntactic unit. All counts reflect CSS only.

Table 1: 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
Cree	1	2;1.14-3;8.24	25	16,797	40,612
English	12	1;8.22-3;0.2	804	373,934	1,443,404
Indonesian	8	1;6.15-8;9.29	997	437,303	1,242,914
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

5. Results

5.1. Comparison of parameters

First we compared different parameter settings for the various surface-level automatic extraction algorithms to get an overall idea of to what extent they affect the identification of variation sets in, and across, languages.[5] We looked at utterance window length from 2-10 and compared the anchor versus incremental approaches.

We find that variation sets in CDS with both methods are ubiquitous in the CSS across languages in our sample. When we apply the anchor method, we find that in all but one language (Inuktitut, which constitutes one of the smallest samples in our corpus), the utterances that belong to variation sets make up proportions greater than 40% in the youngest age ranges (i.e., less than 2;3.0). Even larger proportions were observed for the incremental method. This discrepancy was expected given the sometimes large distances between comparisons that occur in the anchor method (up to eight intervening parental utterances in window size 10), compared to those which occur under the incremental approach (always zero intervening parental utterances). Beyond the general differences, both approaches yield similar patterns of development over time per language. Of particular note is the fact that not all languages exhibit a diminution in the proportion of utterances in variation sets over time, a point to which we return below. We also find that variation sets in CSS with both methods are ubiquitous with nouns and verbs as the main exponents of variation sets as stated in Waterfall (2006).[6]

5. We present our results in detailed reports online at: https://github.com/acqdiv/.

6. We applied the same approach to morphemes instead of words. We found nearly identical results, suggesting that both levels of analysis provide convergent evidence of the same underlying pattern, irrespective of the particular language. However, we

In summary, when we repeat the analysis across all window sizes and numbers of lexical matches, we see that the trends of variation sets are stable across these parameters. Furthermore, the results are similar regardless of anchor versus incremental pairwise comparison, word- versus morpheme-based analysis, or whether we consider all words or only nouns and verbs. The results within languages are also consistent across parameter settings. Variation sets as identified through automatic surface-level extraction algorithms are ubiquitous and robust across the languages in our sample.

5.2. Function of age

Whereas other studies (Waterfall et al., 2010; Wirén et al., 2016) report decreases in the proportion of utterances that belong to variation sets in CDS, we find that this only holds for some of the languages in our sample. Our results therefore corroborate the findings of Grigonyte & Bjorkenstam (2016), who also report that they did not observe the expected decrease across the board in their sample of 26 languages. However, a novel aspect of our findings is that in some cases, the rates may actually increase over time for certain languages. For example, while in Russian the decrease is clear, in Chintang, a polysynthetic language spoken in Nepal, we observe an increase in the proportion of variation sets up to roughly the age of 5.

To test whether variation sets are a function of age, we performed a linear mixed-effect regression predicting (logit-transformed)[7] proportion of variation sets in the child-directed speech of nine languages. For this preliminary analysis, we only consider matches based on nouns or verbs using the anchor method. Our model compares the effects of age across window size (2-10 utterances) and number of matches (1-4 matches). We further include a main effect of language, two three-way interaction terms between language, age and each of the other two predictors, respectively, and all subordinate two-way interactions. Age, number of matches, and window size were centered using *z*-scores. Random intercepts were included for speaker (nested into language) and recording session. All predictors and interactions reached significance at α=.05, except for the two-way interaction between age and window size (meaning that altering window size did not produce reliably different proportions of utterances in variation sets across the ages sampled here when the effect of language is factored out). In English,

---

acknowledge several outstanding problems of morphological analysis: syncretism (where identical forms serve different purposes within the same paradigm), homophony (where identical forms apply to multiple word classes, e.g., English verbal -s 3SG.PRS.IND vs. nominal -s PL), and variable ordering within the same stem (e.g., the affix movement in Chintang verbal complexes (Bickel et al., 2017). Each of these issues could lead to spurious or non-recovered matches.

7. This transformation was necessary given the bounded nature of the response variable. Whereas proportions must fall in the range {0,1}, linear regression effectively predicts values outside of this range at the margins.

Japanese, Russian, Sesotho, and Yucatec, we observe the expected decrease in proportion of utterances belonging to variation sets in child-surrounding speech over time. However, in Chintang, Cree, Inuktitut, and Turkish, the percentage increases as the children grow older (word-level analysis, anchor method, strict matching). We exemplify these trends with Chintang (Figure 1) and Russian (Figure 2), respectively, though several other languages also show this contrast. Thus, there is no one-way trend in our dataset: the number and proportion of variation sets in the input to children differ across languages and with respect to whether or not they decrease as a function of age (Waterfall et al., 2010; Wirén et al., 2016).

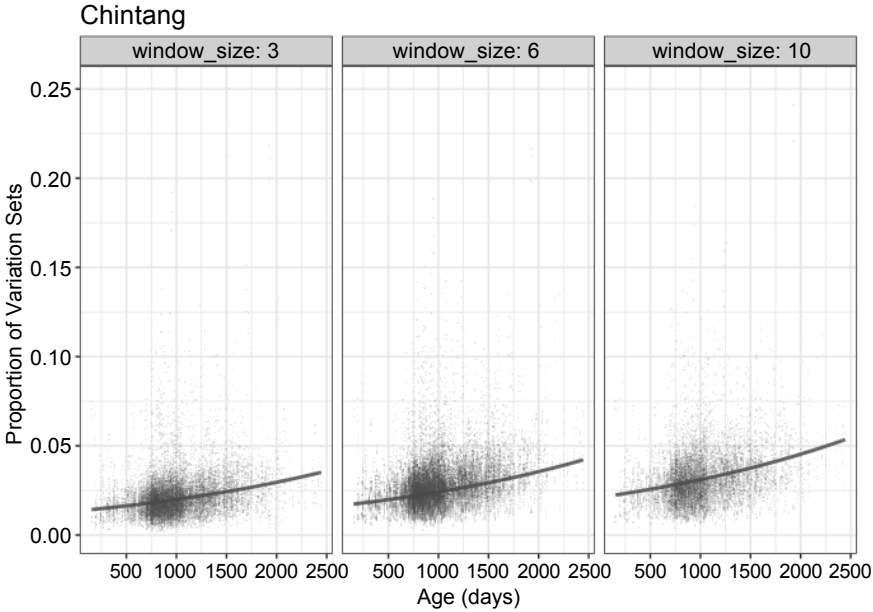

Figure 1: Proportion of variation sets in Chintang over time

6. Discussion

Our results show that variation sets are ubiquitous in child surrounding speech in typologically different languages. Prior research has reported predominantly decreasing trends in the proportion of utterances that belong to variation sets over time; however, at least one study has reported rates that do not decrease over time. We replicate this finding, but go further to show that variation sets may actually increase as a function of age. In our sample, the proportion of variation sets in the input to children tends to increase in the languages with complex morphology (e.g., Chintang, Cree, Inuktitut, and Turkish), but decrease in those with less complex morphology (e.g., English, Japanese, Russian, and Sesotho).

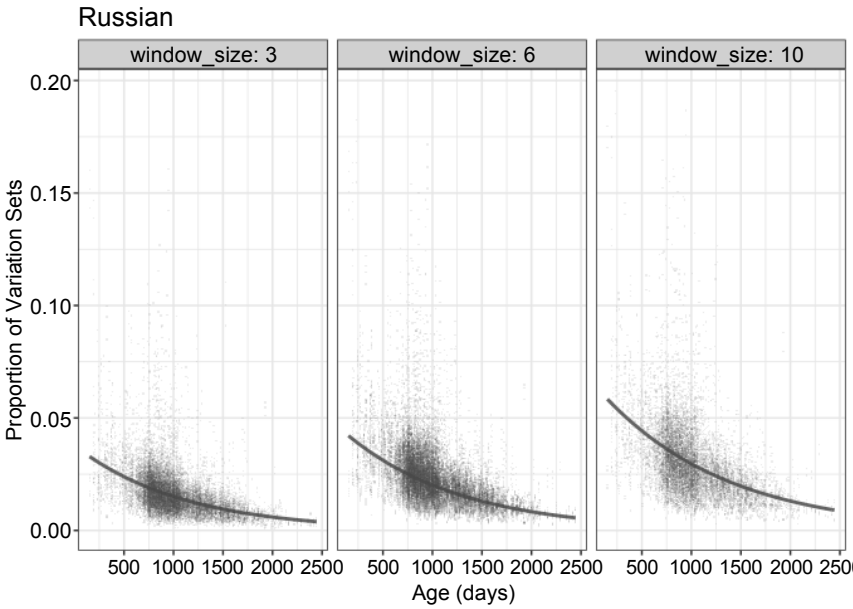

Figure 2: Proportion of variation sets in Russian over time

Thus, the proportion of variation sets seems to be influenced by variables other than age. This observation suggests that we might want to evaluate the extent to which variation sets are a function of morphological complexity and typological characteristics. This can be tested by comparing the proportion of variation sets in adult-to-adult speech of languages with different morphological complexity to test whether morphologically complex languages tend to have a higher baseline proportion of variation sets. In this way, we can determine whether the present findings are merely an artefact of the typological differences between languages, or if they reflect a deeper relationship between typology and child-surrounding speech. That is, languages, given the nature of their morphology, may be more or less repetitive. Therefore, it could be that our approaches for identifying variation sets, particularly at the morpheme level, are simply picking up on these aspects of grammar.

Another point to consider is the possible effect of morphology on the change in prevalence of variation sets over time. Languages with shallow morphology, such as English, provide fewer opportunities for overlap across words with different roots. As speech becomes more lexically diverse, the likelihood of identifying a variation set necessarily decreases. The observed decrease for such languages suggests that adults indeed begin to produce more lexically variable utterances as the child grows, perhaps due to the fact that the communicative demands of adult-child interaction become more complex. By contrast, in morphologically complex languages, the inflectional components of words provide much more overlap between the inflected forms of different stems. For

languages with small roots and numerous morphological slots (e.g., Chintang), we might observe increasing proportions of utterances identified as variation sets as speech becomes more lexically diverse with respect to the root. Another possibility is that adults slowly increase the morphological complexity and variability of their utterances as children become more proficient speakers, which would in turn create more possibility for overlap between words and morphemes across utterances. Further research is needed to test these possibilities.

Other variables could also 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 (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 vs. lower SES language samples from English and Hebrew have statistically significantly higher proportions of variation sets. Do these results hold when the parameters that identify variation sets are shifted?

A shortcoming of extraction techniques as presented in the present paper is that they are based on surface forms and so cannot address suppletive semantic repetitiveness (Grigonytė & Björkenstam, 2016). Therefore, another area for future research is to move beyond surface-based automatic extraction techniques. Ideally, we need to move away from surface-based approaches and towards semantically-informed methods, such as Latent Semantic Analysis (Landauer & Dumais, 1997) and Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). These approaches, however, typically need a lot of training data, and as such, are not feasible for the under-resourced languages in our sample. Beyond collecting more data, semantic similarity approaches that can work with smaller datasets should be explored.

The data that we analyze statistically also represent only a small fraction of the possible ways that the variation sets can be extracted. A more detailed analysis would consider what impact the different parameters have on the shape of the observed effects once entered into a carefully controlled statistical model. Nevertheless, the current findings provide important information about how the decisions we make impact the behavior of models that are based on automatically detected variation sets, as well as how these outcomes differ between languages.

References

Allen, Shanley. (unpublished). Allen Inuktitut Child Language Corpus.

Bard, Ellen G., & Anderson, Anne H. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10*(2), 265–292.

Björkenstam, Kristina N., & Wirén, Mats. (2014). Multimodal annotation of synchrony in longitudinal parent–child interaction. In *The 9th Edition of the Language Resources and Evaluation Conference (LREC), 26-31 May, Reykjavik, Iceland*. European Language Resources Association.

Brittain, Julie. (2015). Corpus of the Chisasibi Child Language Acquisition Study (CCLAS). Retrieved from http://phonbank.talkbank.org/access/Other/Cree/CCLAS.html

Brodsky, Peter, Waterfall, Heidi, & Edelman, Shimon. (2007). Characterizing motherese: On the computational structure of child-directed language. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 29).

Cameron-Faulkner, Thea, Lieven, Elena, & Tomasello, Michael. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843c873.

Demuth, Katherine. (1992). Acquisition of Sesotho. In D. Slobin (Ed.), The Cross-Linguistic Study of Language Acquisition, Vol. 3 (pp. 557-638). Hillsdale: Lawrence Erlbaum Associates.

Fernald, Anne, & Hurtado, Nereyda. (2006). Names in frames: infants interpret words in sentence frames faster than words in isolation. *Developmental Science*, *9*(3), F33–F40.

Frank, Michael C., Tenenbaum, Joshua B., & Fernald, Anne. (2013). Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning. *Language Learning and Development: The Official Journal of the Society for Language Development*, *9*(1), 1–24.

Gil, David, & Tadmor, Uri. (2007). The MPI-EVA Jakarta Child Language Database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.

Grigonytė, Gintarė, & Björkenstam, Kristina 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* (pp. 41–50). Linköping University Electronic Press.

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

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

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

Küntay, Aylin C., Koçbaş, Dilara, & Taşçı, Süleyman S. (2015). KULLDD, Koç University Longitudinal Language Development Database on language acquisition of 8 children from 8 to 36 months of age.

Küntay, Aylin C., & Slobin, Dan I. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In Dan I. Slobin, Julie Gerhardt, Amy Kyratzis, and Jiangsheng Guo (Ed.), *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp* (pp. 265–286). Lawrence Erlbaum Associates, Inc.

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

Landauer, Thomas K., & Dumais, Susan T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

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

Lew-Williams, Casey, Pelucchi, Bruna, & Saffran, Jenny R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329.

MacWhinney, Brian. (2000). The CHILDES project: Tools for analyzing talk. *Mahwah, NJ*.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

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

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

Moran, Steven, Blasi, Damián E., Schikowski, Robert, Küntay, Aylin C., Pfeiler, Barbara, Allen, Shanley, & Stoll, Sabine. (2018). A universal cue for grammatical categories in the input to children: Frequent frames. *Cognition*, *175*, 131–140.

Moran, Steven, Schikowski, Robert, Pajović, Danica, Hysi, Cazim, & Stoll, Sabine. (2016). The ACQDIV Database: Min (d) ing the Ambient Language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* 23—28 May, Portorož, Slovenia.

Onnis, Luca, Waterfall, Heidi, & Edelman, Shimon. (2008). Learn locally, act globally: learning language from variation set cues. *Cognition*, *109*(3), 423–430.

Pfeiler, Barbara. (unpublished). Pfeiler Yucatec Child Language Corpus.

Pfeiler, Barbara. (2017). The acquisition of action nouns in Yucatec Maya. In Valentina Vapnarsky & Edy Veneziano (Eds.), *Lexical Polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (pp. 443-466). Amsterdam: John Benjamins.

Ratcliff, John W., & Metzener, David E. (1988). Pattern-matching-the gestalt approach. *Dr. Dobb's Journal*, *13*(7), 46.

Redington, Martin, Chater, Nick, & Finch, Steven. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science*, *22*(4), 425–469.

Rowland, Caroline F., & Fletcher, Sarah L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language*, *33*(4), 859–877.

Rowland, Caroline F., Pine, Julian M., Lieven, Elena, & Theakston, Anna L. (2005). The incidence of error in young children's Wh-questions. *Journal of Speech, Language, and Hearing Research*, *48*(2), 384–404.

Stoll, Sabine, Abbot-Smith, Kirsten, & Lieven, Elena. (2009). Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech. *Cognitive Science*, *33*(1), 75–103.

Stoll, Sabine, & Bickel, Balthasar. (2013). Capturing diversity in language acquisition research. *Language Typology and Historical Contingency: In Honor of Johanna Nichols*, 195–216. Amsterdam: Benjamins.

Stoll, Sabine, Bickel, Balthasar, Lieven, Elena, Paudyal, Netra P., Banjade, Goma, Bhatta, Toya N., Gaenszle, Martin, Pettigrew, Judith, Rai, Ichchha Purna, Rai, Manoj, Rai, Novel Kishore. (2012). Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language*, *39*(2), 284–321.

Stoll, Sabine, Lieven, Elena, Banjade, G., Bhatta, Toya N., Gaenszle, Martin, Paudyal, Netra P., Rai, Manoj, Rai, Novel Kishore, Rai, Ichchha Purna, Zakharko, Taras,

Schikowski, Robert, & Bickel, Balthasar. (unpublished). Audiovisual corpus on the acquisition of Chintang by six children.

Stoll, Sabine, & Meyer, Roland. (2008). Audio-visual longitudinal corpus on the acquisition of Russian by 5 children.

Tal, Shira, Arnon, Inbal. (2018). SES Differences in the Communicative Functions of Variation Sets. In Anne B. Bertolin & Maxwell J. Kaplan (Eds.), *BUCLD 42: Proceedings of the 42nd Annual Boston University Conference on Language Development* (pp. 736–749). Cascadilla.

Theakston, Anna L., Lieven, Elena, Pine, Julian M., & Rowland, Caroline F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, *28*(1), 127–152.

Waterfall, Heidi. (2006). A little change is a good thing: Feature theory, language acquisition and variation sets. *Language Acquisition and Variation Sets. Chicago: University of Chicago*.

Waterfall, Heidi, Sandbank, Ben, Onnis, Luca, & Edelman, Shimon. (2010). An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language*, *37*(3), 671–703.

Wirén, Mats, Björkenstam, Kristina, N., Grigonytė, Gintar ė, & 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.

Proceedings of the 43rd annual Boston University Conference on Language Development

edited by Megan M. Brown
and Brady Dailey

Cascadilla Press Somerville, MA 2019

ISSN 1080-692X
ISBN 978-1-57473-096-8 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com