Language-independent exploration of repetition and variation in longitudinal child-directed speech: a tool and resources

Gintarė Grigonytė and Kristina Nilsson Björkenstam
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
gintare@ling.su.se, kristina.nilsson@ling.su.se

Abstract

We present a language-independent tool, called *Varseta*, for extracting variation sets in child-directed speech. This tool is evaluated against a gold standard corpus annotated with variation sets, MINGLE-3-VS, and used to explore variation sets in 26 languages[1] in CHILDES-26-VS, a comparable corpus derived from the CHILDES database. The tool and the resources are freely available for research.[2]

1 Introduction

Repetitiousness is a strong trait of child-directed speech. When parents speak to young infants, a large proportion of utterances are either exact repetitions of an immediately preceding utterance, or partial repetitions, where the message is repeated and thus, the speaker intent is constant, but variation occurs in the surface form. Such sequences of partial repetitions were first referred to as *variation sets* by Küntay and Slobin (1996). Surface form variation includes expansion, insertion, deletion, and word order change, e.g.:

le petit chat? ('the small cat?')[3]
tu m'aides? ('will you help me?')
tu m'aides à chercher? ('will you help me look?')
il est où là le petit chat? ('where is the small cat?')

The repetitiousness can also be semantic, e.g., in cases of lexical substitution such as this where the verbs *titta*, *sett*, *kolla* are variations of 'to look (at something)' (Wirén et al., 2016):

titta här då! ('look at this!')[4]
har du sett vilka tjusiga byxor? ('have you seen such fancy pants?')
kolla! ('check it out!')

Current research suggests that such sequences of repetition and variation play a role in language learning, e.g., experiments on artificial language learning and variation sets (Onnis et al., 2008), as well as child corpus studies on correlations between variation sets and language acquisition (Hoff-Ginsberg, 1986; Hoff-Ginsberg, 1990; Waterfall, 2006; Küntay and Slobin, 1996). This paper builds upon these assumptions, but does not concern the output of the learner. Rather, our aim is to investigate the input to the learner, and more specifically, the longitudinal patterns of occurrences of variation sets in child-directed speech across multiple languages.

To our knowledge, variation sets have been studied in Turkish (Küntay and Slobin, 1996; Küntay and Slobin, 2002), English (Waterfall, 2006), Sign Language of the Netherlands (Hoiting and Slobin,

[1] 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.

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

[3] Example from CHILDES FrenchGeneva14.cha, PID: 11312c-00028164-1. English translations are approximate.

[4] Example from (Wirén et al., 2016). English translations are approximate.

Gintarė Grigonytė and Kristina Nilsson Björkenstam 2016. Language-independent exploration of repetition and variation in longitudinal child-directed speech: A tool and resources. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*. Linköping Electronic Conference Proceedings 130: 41–50.

2002), and Swedish, English, Russian, and Croatian (Wirén et al., 2016). Studies using longitudinal data have shown that as the communication skills of the child increase, the proportion of utterances in variation sets decreases (Waterfall et al., 2010; Wirén et al., 2016).

This study expands the scope of previous work by using a large-scale cross-language approach to explore repetition and variation in child-directed speech. Further, the approach proposed in this paper on extracting variation sets from transcripts of child-directed speech is language-independent and automatic. This paper presents two surface-based strategies for automatic variation detection (see section 4). The strategies are evaluated against a gold standard corpus annotated according to the annotation scheme for variation sets described in (Wirén et al., 2016).

2 Related work

While most definitions of variation sets include both speaker intention and utterance form (c.f., (Küntay and Slobin, 1996; Küntay and Slobin, 2002; Waterfall, 2006; Wirén et al., 2016)), previous attempts at automatic extraction of variation sets focus primarily on form.

Brodsky et al. (2007) suggest a narrower definition of variation set as sequences of utterances where each successive pair of utterances has a lexical overlap of at least one element. Variation sets can thus be extracted by comparing pairs of successive utterances for repeated words, resulting in sets with at least one word in common. Using such an extraction procedure, Brodsky et al. found that 21.5% of the words in Waterfall's (2006) corpus (12 mother–child dyads, child age 1;2-2;6 years) occur in variation sets, and 18.3% of the words in the English CHILDES database (MacWhinney, 2000).

Similarly, Onnis et al. based their extraction strategy on Waterfall's (2006) criteria for variation sets. When applied to the CHILDES Lara corpus (child age 1;9–3;3 years), 27,9% of the utterances were extracted as belonging to variation sets.

Also using a surface-based algorithm for automatic extraction of variation sets, but with a novel definition of variation sets, Wirén et al. (2016) show that the proportion of variation sets in child-directed speech decreases consistently as a function of children's age across Swedish, Croatian, English and Russian. They report fuzzy F-scores of 0.822, 0.689, 0.601, and 0.425 for 4 age groups in Swedish data respectively.

This study expands the scope of the latter paper in two ways: a) by offering two variation set extraction strategies ANCHOR and INCREMENTAL which are evaluated against a gold standard corpus of Swedish; b) by using these strategies in a large-scale cross-language investigation of child-directed speech corpora derived from the CHILDES database (MacWhinney, 2000); c) by releasing the software and the derived corpora along with the gold standard corpus of Swedish.[5]

3 Data sets

We use two different data sets for exploration of repetition and variation in child-directed speech. The longitudinal Swedish corpus, MINGLE-3-VS, is annotated with variation sets. The second data set, here called CHILDES-26-VS, consists of plain text transcripts of child-directed speech in 26 languages, derived from the CHILDES database (MacWhinney, 2000). The corpus files are grouped by language and child age which allows for both cross-language and within-language longitudinal comparisons.

3.1 MINGLE-3-VS: a corpus annotated with variation sets

The gold standard variation set corpus, MINGLE-3-VS, consist of transcripts of Swedish child-directed speech annotated with variation sets according to the annotation scheme described in (Wirén et al., 2016).

The transcripts originates from the MINGLE-3 multimodal corpus (Björkenstam et al., 2016), which consists of 18 longitudinal dyads with three children (two girls, one boy; six dyads per child) recorded between the ages of 7 and 33 months. The complete duration of the 18 dyads is 7:29 hours (mean duration 24:58 minutes). The video and audio recordings were made from naturalistic parent–child interaction in a studio at the Phonetics Laboratory at Stockholm University (Lacerda, 2009). The children were interacting alternately with their mothers (10 dyads) and fathers (8 dyads) in a free play sce-

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

CHILDES language group	Language	Corpora	# Children	Age span	# Dyads
Celtic	Irish	Gaeltacht	1	3,4	2
	Welsh	CIG1	1	3,4	2
EastAsian	Cantonese	HKU, LeeWongLeung	2	3,4	3
	Chinese	Beijing, XuMinChen, Zhou1	7	2,3,4	7
	Indonesian	Jakarta	2	3,4	2
	Japanese	Ishii, Miyata	2	1,2,3,4	6
	Thai	CRSLP	1	1,2,3,4	4
Germanic	Afrikaans	VanDulm	2	3,4	4
	Danish	Plunkett	2	1,2,3,4	5
	Dutch	Groningen, VanKampen	2	3,4	4
	English UK	Lara	1	3,4	2
	German	Caroline, Manuela, Szagun	4	1,2,3,4	9
Romance	Catalan	Julia	1	3,4	2
	French	Geneva, Hunkeler, Lyon, Pauline	4	1,2,3,4	8
	Italian	Antelmi, Calambrone	2	3,4	3
	Portuguese	Santos	2	3,4	2
	Spanish	Irene	1	1,2,3,4	4
Slavic	Croatian	Kovacevic	1	1,2,3	3
	Russian	Protassova	1	3,4	2
Other	Estonian	Argus, Kapanen, Kohler, Zupping	5	1,2,3,4	7
	Farsi	Samadi	2	3,4	2
	Greek	Stephany	1	3,4	2
	Hebrew	BSF, Levy, Naama	4	2,3,4	4
	Hungarian	Bodor, MacWhinney, Reger	3	3,4	3
	Tamil	Narasimhan	1	1,2,3,4	4
	Turkish	Aksu, Turkay	2	3,4	4
Total	26 languages	45 corpora	60	–	100

Table 1: CHILDES-26-VS: Corpora derived from the CHILDES database (MacWhinney, 2000) grouped by CHILDES language group, and presented per language. Col. 3: corpus name(s), col. 4: total number of children, col. 5: the age groups covered (1–4), col. 6: the total number of dyads.

nario.[6] The ELAN annotation tool (Wittenburg et al., 2006) was used for transcription of parent and child utterances, as well as non-verbal annotation (Björkenstam et al., 2016).

ELAN was also used for manual variation set annotation. This allowed for the annotators to take both verbal and non-verbal input from parent and child into account when deciding on the boundaries of variation sets. The annotation methodology was as follows: during the first phase, a subset of four dyads was annotated by two coders independently. After merging the respective annotations for each dyad, a third annotator marked cases of disagreement. This resulted in an inter-annotator agreement (measured as set overlap between annotators) of 78%. Disagreements were solved during group discussions. After evaluation of the first phase, the remaining 14 dyads were annotated by one annotator. Finally, a classification of communicative intention based on the Inventory of Communicative Acts-Abridged (Ninio et al., 1994) was added. This classification was evaluated by comparing four representative dyads annotated by three independent annotators, resulting in a Fleiss's kappa of 0.63. The transcripts were also annotated with part-of-speech using Stagger (Östling, 2013), followed by manual

[6] A subset of the audio files is available through CHILDES/Swedish/Lacerda (MacWhinney, 2000).

correction (Wirén et al., 2016).

3.2 CHILDES-26-VS: corpora derived from CHILDES

We have extracted child-directed speech from transcripts in 45 corpora in 26 languages from the CHILDES database (MacWhinney, 2000). The selection criteria was the scenario (naturalistic interaction), the participants (parents or other adults - including researchers - and children), and the age of the child (0;6 to 2;9 years). The selected transcripts were grouped according to child age. The grouping approximates major physical child development stages, i.e., sitting up (0;6–0;11 years), standing-walking (1;0–1;3 years), fully mobile (1;4–1;11 years), and talking (2;0–2;9 years) (see table 2).

	MINGLE-3-VS	CHILDES-26-VS
Age group 1	0;6 – 0;9	0;6 – 0;11
Age group 2	1;0 – 1;2	1;0 – 1;3
Age group 3	1;4 – 1;7	1;4 – 1;11
Age group 4	2;3 – 2;9	2;0 – 2;9

Table 2: Age groups and age spans (year;months) in MINGLE-3-VS and the derived corpora CHILDES-26-VS.

An overview of the sources is presented in table 1, detailing for each language the language group according to CHILDES, the name of the corpus or corpora, the total number of children, the age groups (1–4) covered by the transcripts, and the total number of transcripts.

All files in the derived corpora are grouped by language and child age which allows for both cross-language and within-language longitudinal comparisons of variation sets with the Varseta tool. This data set is freely available for research as part of the Varseta package (see section 4).

4 Varseta - a tool for automatic extraction of variation sets

The Varseta tool for variation set extraction for any language is available at GitHub[7].

The definition of variation sets that we follow in the implementation of the Varseta tool takes into account exact repetitions, and further allows the following transformations between utterances: reduction, expansion, and word order change (Wirén et al., 2016). Although these alternations might be fairly complex, a large proportion of them can be observed on the surface level, and thus automatically extracted on the basis of string similarity techniques.

Varseta employs two commonly used string similarity measures: the Ratcliff-Obershelp pattern recognition method (Black, 2004) and edit distance ratio[8] (Levenshtein, 1966), and uses two strategies for detecting variation sets in child-directed speech: ANCHOR and INCREMENTAL. The two string similarity measures and the two strategies can be used in any combination, allowing for 4 different settings.

For a given set of utterances, the ANCHOR strategy measures pairwise utterance similarity of all utterances in relation to the first, e.g. 1-2, 1-3, 1-4. The criterion for including two utterances in a variation set is that the difference between them (regarded as strings) does not fall below a certain similarity threshold. Additionally, following Brodsky et al. (2007), we allow for sequences of maximally two intervening dissimilar utterances that do not obey this condition.

For a given set of utterances, the INCREMENTAL strategy performs a stepwise comparison of pairs of successive utterances, e.g. 1-2, 2-3, 3-4. Two utterance strings that pass a certain similarity threshold are marked as belonging to a variation set. Unlike the ANCHOR strategy, sequences of intervening dissimilar utterances are not allowed. Thus the process continues, by adding similar utterances, until a non-similar utterance occurs.

Both strategies can employ either edit distance ratio (**EDR**) or Ratcliff-Obershelp pattern recognition method (**DLR**, as implemented in the Python module difflib[9]). String similarity measures return values between [0..1], convenient for categorizing string utterances on the surface level. A value of 1 means exact repetition of an utterance, and 0 means two unrelated utterances without any overlap of words. The similarity threshold used in this experimental study, as described in section 5, was arbitrarily selected. The most optimal similarity thresholds when evaluated against the Swedish gold stan-

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

[8] Also known as Levenshtein distance.

[9] difflib: https://docs.python.org/2/library/difflib.html#module-difflib

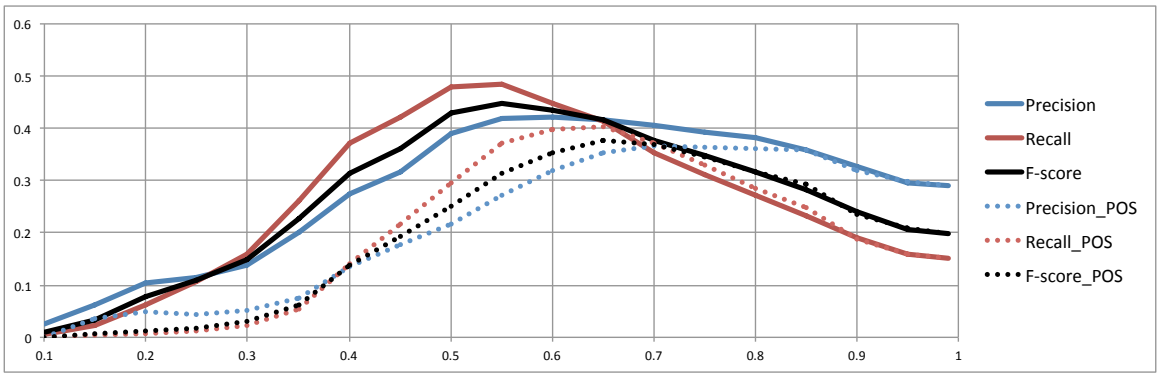

Figure 1: Results of Varseta strict matching with ANCHOR and the DLR similarity measure on raw (solid lines) and part-of-speech tagged data (dotted lines). Similarity level thresholds on x-axis; precision, recall and F-score on y-axis.

dard was 0.55 for DLR, and 0.51 for EDR (see Figure 1).

While performing experiments on the Swedish gold standard data, we found that the ANCHOR strategy with the DLR similarity measure performed slightly better relative to the gold standard annotation.

Additionally, we experimented on including information from the part-of-speech tagging of the transcripts in such away that the pair of strings compared consisted of both the words and their part-of-speech tags. Our intuition was that this might give a more refined analysis, for example, by distinguishing cases of homonymy. This version of the algorithm turned out not to improve performance, however (see Figure 1), and was therefore dropped.

5 Results: Automated extraction of variation sets

5.1 Evaluation against the Swedish gold standard

We evaluated the Varseta tool against the gold standard using two kinds of metrics, which we refer to as *strict* and *fuzzy* matching. Strict matching requires exact matching on the utterance level of the extracted variation set and the corresponding gold standard set, whereas fuzzy matching allows for partial overlaps of the extracted variation set and the gold standard set. In the example in Table 3, only utterance 3 and 4 are members of the gold standard variation set, whereas the algorithm extracts utterances 1–4. Hence, the strict matching metric treats this extracted set as a false positive, whereas the fuzzy matching metric treats it as a true positive.

Table 4 summarizes the results of extraction of variation sets relative to the gold standard according to the strict and fuzzy metric. Strict F-score reaches 0.577 and fuzzy F-score reaches 0.813 for age group 1, but F-scores gradually decrease with increasing age.

This observed phenomenon has two reasons: first, the decrease in the proportion of exact repetitions as the child grows older; second, the increasing complexity of the parent's speech. As the complexity increases, capturing variation requires more than surface-based methods. This finding is in line with (Wirén et al., 2016).

5.2 Extraction of variation sets in 26 languages

For exploration of repetition and variation in child-directed speech in 26 languages, as captured in CHILDES-26-VS, we have used the Varseta tool.

We expected to find decreasing **proportions of utterances in variation sets** as a function of child age for all languages.

The findings for a majority of languages, 19 out of 26 (Irish, Welsh, Cantonese, Indonesian, Japanese, Afrikaans, Danish, Dutch, English, German, Swedish, Italian, Spanish, Croatian, Russian, Estonian, Farsi, Greek, and Turkish), indicate a decrease in the proportion of utterances in variation sets as a function of child age (see bold face proportions in table 5 on page 7).

We have observed exceptions in Chinese, Thai, Catalan, French, Portuguese, Hebrew, and Tamil.

Example utterances	Member of gold set	Extracted by algorithm
1. *Ska vi lägga ner nånting i i väskan då?* (‘Are we going to put something in in the bag then?’)	–	Yes
2. *Va?* (‘Huh?’)	–	Yes
3. *Ska du lägga ner kossan i väskan kanske?* (‘Are you going to put down the cow in the bag maybe?’)	Yes	Yes
4. *Ska vi lägga ner kossan?* (‘Are we going to put down the cow?’)	Yes	Yes

Table 3: Example variation set from the gold standard (utterance 3–4) and utterances extracted by the Varseta tool (utterance 1–4).

String matching relative to gold standard	Group 1 0;7–0;9	Group 2 1;0–1;2	Group 3 1;4–1;7	Group 4 2;3–2;9
ANCHOR				
Strict precision	0.554	0.415	0.337	0.164
Strict recall	0.603	0.460	0.473	0.282
Strict *F*-score	**0.577**	0.437	0.393	0.208
INCREMENTAL				
Strict precision	0.549	0.476	0.416	0.415
Strict recall	0.559	0.418	0.453	0.436
Strict *F*-score	0.554	**0.445**	**0.433**	**0.425**
ANCHOR				
Fuzzy precision	0.779	0.634	0.548	0.358
Fuzzy recall	0.849	0.703	0.770	0.615
Fuzzy *F*-score	**0.813**	**0.667**	**0.640**	0.453
INCREMENTAL				
Fuzzy precision	0.736	0.621	0.553	0.537
Fuzzy recall	0.748	0.545	0.601	0.564
Fuzzy *F*-score	0.742	0.581	0.576	**0.550**

Table 4: Evaluation of the Varseta tool for automatic variation-set extraction against the Swedish gold standard per age group.

For Chinese, Thai, Hebrew, and Tamil, there are insufficient amounts of data for earlier age groups (age groups 2, 1, 2, and 1, respectively) which skews the proportion in comparison to older age groups. For instance, Chinese age group 2 contains 294 utterances and Chinese age group 3 contains 1395. In the age groups with sufficient/comparable amounts of data for these three languages, we do observe the expected decrease pattern.

However, data insufficiency or incomparability cannot explain the unexpected findings for French and Portugese, and thus in-depth analysis of these transcripts is needed.

For most of the languages the similar pattern of decrease in **proportion of exact repetitions** cannot be observed. One general trend is that the proportion of exact repetitions is small as compared to the proportion of utterances in variation sets. Exceptions to this trend are observed in Swedish, Danish, English, Russian, Cantonese, Japanese, Thai, Welsh, Estonian, Hebrew, and Tamil (average proportion of exact repetition: 0.13 in age group 1, 0.096 in age group 2, 0.066 in age group 3, and 0.04 in age group 4).

For some languages we observe a decrease in proportions, even to no exact repetitions, for example in German, French, Italian, Spanish, Farsi, and Turkish. A close inspection of these data files revealed

Lang. group	Language		Age groups				Lang. group	Language		Age groups			
			1	**2**	**3**	**4**				**1**	**2**	**3**	**4**
Celtic	Irish	a	–	–	862	899	Romance	Catalan	a	–	–	47	264
		b	–	–	**0.63**	**0.39**			*b*	–	–	*0.45*	*0.61*
		c	–	–	0.03	0.05			*c*	–	–	0.09	0.02
	Welsh	a	–	–	304	226		French	a	420	281	450	308
		b	–	–	**0.54**	**0.23**			*b*	*0.38*	*0.49*	*0.41*	*0.49*
		c	–	–	0.11	0.04			*c*	0.00	0.05	0.04	0.06
EastAsian	Cantonese	a	–	–	392	1278		Italian	a	–	–	368	541
		b	–	–	**0.65**	**0.58**			*b*	–	–	**0.46**	**0.37**
		c	–	–	0.07	0.04			*c*	–	–	0.01	0.00
	Chinese	a	–	294	1395	1338		Portuguese	a	–	–	783	660
		b	–	*0.57*	**0.64**	**0.57**			*b*	–	–	*0.57*	*0.61*
		c	–	0.02	0.04	0.02			*c*	–	–	0.04	0.01
	Indonesian	a	–	–	577	714		Spanish	a	81	44	122	221
		b	–	–	**0.58**	**0.50**			*b*	**0.70**	**0.57**	**0.45**	**0.25**
		c	–	–	0.05	0.05			*c*	0.05	0.00	0.00	0.00
	Japanese	a	220	281	525	1315	Slavic	Croatian	a	39	217	408	–
		b	**0.91**	**0.79**	**0.62**	**0.60**			*b*	**0.85**	**0.54**	**0.50**	–
		c	0.17	0.10	0.04	0.06			*c*	0.00	0.09	0.05	–
	Thai	a	123	222	172	250		Russian	a	–	–	1088	545
		b	*0.42*	*0.50*	*0.49*	*0.51*			*b*	–	–	**0.35**	**0.24**
		c	0.07	0.03	0.03	0.02			*c*	–	–	0.06	0.05
Germanic	Afrikaans	a	–	–	87	128	Other	Estonian	a	58	527	420	383
		b	–	–	**0.56**	**0.54**			*b*	**0.62**	**0.37**	0.43	0.41
		c	–	–	0.09	0.00			*c*	0.10	0.03	0.02	0.02
	Danish	a	136	630	250	582		Farsi	a	–	–	103	32
		b	**0.82**	**0.65**	**0.67**	**0.53**			*b*	–	–	**0.64**	**0.41**
		c	0.23	0.13	0.10	0.05			*c*	–	–	0.00	0.00
	Dutch	a	–	–	989	1176		Greek	a	–	–	246	453
		b	–	–	**0.52**	**0.50**			*b*	–	–	**0.56**	**0.48**
		c	–	–	0.06	0.03			*c*	–	–	0.06	0.07
	English	a	–	–	926	391		Hebrew	a	–	132	156	108
		b	–	–	**0.54**	**0.44**			*b*	–	*0.51*	*0.68*	*0.65*
		c	–	–	0.08	0.07			*c*	–	0.08	0.08	0.04
	German	a	82	62	1160	586		Tamil	a	54	239	220	182
		b	**0.77**	**0.55**	**0.51**	0.54			*b*	*0.65*	*0.82*	*0.68*	*0.70*
		c	0.07	0.00	0.04	0.10			*c*	0.04	0.08	0.07	0.04
	Swedish[10]	a	1032	1421	1483	724		Turkish	a	–	–	567	322
		b	**0.61**	**0.43**	**0.52**	**0.36**			*b*	–	–	**0.50**	**0.46**
		c	0.18	0.08	0.07	0.04			*c*	–	–	0.01	0.01

Table 5: Results of the ANCHOR-DLR strategy for automatic variation-set extraction applied to CHILDES-26-VS. Results are grouped by CHILDES language group (col. 1) and language (col. 2). For each language, a) the number of utterances, b) the proportion of utterances in variation sets, and c) the proportion of exact repetitions per age groups 1 (0;6–0;11), 2 (1;0–1;3), 3 (1;4–1;11), and 4 (2;0–2;9). Proportions in bold face follow expectations, whereas proportions in italics do not.

that this is not only the effect of the absence of exact repetitions, but also due to the level of analysis added to the transcripts, for instance markup for perceived pause length or prosody, comments in English, etc.

We also note that about half of the transcripts in Cantonese, Chinese, and Japanese were in latin characters, whereas e.g., some of the transcripts in age group 2 are in Chinese characters. Transcripts written in such logographic systems have a more compressed representation on the utterance level, and thus the similarity measure might need an adjustment.

In addition to quantitative trends across the languages, Varseta also provides variation sets for inspection. Here are two examples of automatically extracted variation sets in German and Farsi.

wo sind die anderen flaschen? ('where are the other bottles?')[11]
guck mal da unten bei dem auto. ('look down there by the car.')
da is noch eine flasche. ('there is one bottle.')
da sin die anderen flaschen. ('there are the other bottles.')

ho gorbe chi mige? ('what does the cat say?')[12]
gorbehe ci mige? ('what does the cat say?')
chi mige? ('what does it say?')
mamaoushe chi mige? ('what does ?(the mouse's mother) say?')
mamoushe chi? ('?(the mouse's mother) what?')

6 Discussion

The evaluation of the Varseta tool for Swedish indicates that variation sets are easier to capture for earlier age groups (ANCHOR fuzzy F-score: 0.813, 0.667, 0.640 and 0.453 for age groups 1, 2, 3 and 4). The F-score reflects on the complexity of the input, that is, not only the proportion of exact repetitions, but also patterns of expansion, insertion, deletion, word order change, and lexical substitution over sequences of utterances. Further, the algorithm does not include information on speaker turns as this information is not available in the current version of the corpus, and it is likely that this contributes to the low precision in the later dyads. According to the definition we follow, child vocalizations are allowed within a variation set (c.f., Wirén et al., 2016), but when such a child utterance constitutes a legitimate turn, the variation set should be split in two. Overall, the performance is according to what can be expected from a simple surface-based method. To our knowledge this is the only extraction method that has been evaluated against a manually annotated gold standard and therefore can serve as a baseline method for similar investigations.

The Varseta tool offers both quantitative analysis of repetition and variation in speech transcripts, and output in the form of sequences of utterances from those transcripts that constitute variation sets.

With regards to the analysis of the CHILDES-26-VS with the Varseta tool, the expected decrease in proportion of utterances in variation sets was observed for the majority of languages. The same observation cannot be made for the proportion of exact repetitions. This may be due to differences in transcription, for example regarding utterance segmentation and pause markup, between corpora in CHILDES. For instance, the Varseta tool cannot recognize variation in this example, as within-utterance repetition is not recognized by the tool. The short intervals, here marked by '(.)', may in another corpus constitute segmentation boundaries:

canta lá (.) canta (.) tu sabes? ('sing there (.) sing (.) you know?')[13]
canta com o patinho. ('sing with the duckling.')

The current method does not take into account semantic variation, complete lexical substitution, and other forms of complex variation. The surface-based approach can be improved by adaptating semantic similarity methods like Word2Vec (Mikolov et al.,

[11]Example from CHILDES German/Szagun/NH/Celina/cel10400.cha, PID: 11312/c-00024238-1.

[12]Example from CHILDES Other/Farsi/Samadi/Shahrzad/sha108.cha, PID: 11312/c-00026963-1. English translations are approximate.

[13]Example from CHILDES Romance/Portuguese/Santos/Ines/1-7-6.cha, @PID: 11312/c-00037400-1

2013) as a possible solution for capturing lexical substitutions.

7 Conclusion

This study expands the scope of previous work by using a large-scale cross-language approach to exploring repetition and variation in child-directed speech. Further, the approach proposed in this paper on extracting variation sets from transcripts of child-directed speech is language-independent and automatic. The Varseta tool uses two surface-based strategies for automatic variation set detection which were evaluated against a gold standard corpus MINGLE-3-VS. The software, the gold standard corpus of Swedish, and the comparable corpus of 26 languages derived from CHILDES are freely available for exploration of repetition and variation in child-directed speech.

We have also reported findings on repetition and variation in child-directed speech in 26 languages, as captured in CHILDES-26-VS, using the Varseta tool. We expected to find decreasing proportions of utterances in variation sets as a function of child age for all languages. The findings confirmed this expectation for a majority of languages, except for French and Portuguese.

Acknowledgments

This research is part of the project "Modelling the emergence of linguistic structures in early childhood", funded by the Swedish Research Council (project 2011-675-86010-31). The development of the Varseta package (inkl. data) has been supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, project 821-2013-2003). Thanks to Lisa Tengstrand and Claudia Eklås Tejman for annotation work, and Annika Otsa for extracting and preparing the CHILDES data. Ghazaleh Vafaeian provided the Farsi translation. Finally, we would like to thank the three anonymous reviewers for valuable comments.

References

Kristina Nilsson Björkenstam, Mats Wirén, and Robert Östling. 2016. Modelling the informativeness and timing of non-verbal cues in parent–child interaction. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, August 11, 2016, Association for Computational Linguistics*, pages 82–90, Berlin, Germany.

Paul E. Black. 2004. Ratcliff/Obershelp pattern recognition. *Dictionary of Algorithms and Data Structures*, 17.

Peter Brodsky, Heidi R. Waterfall, and Shimon Edelman. 2007. Characterizing motherese: On the computational structure of child-directed language. In *Proc. 29th Cognitive Science Society Conference*, Nashville, TN.

Erika Hoff-Ginsberg. 1986. Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology*, 22(3):155–163.

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

Nini Hoiting and Dan I. Slobin. 2002. What a deaf child needs to see: Advantages of a natural sign language over a sign system. In R. Schulmeister and H. Reinitzer, editors, *Progress in sign language research. In honor of Siegmund Prillwitz/Fortschritte in der Gebärdensprachforschung. Festschrift für Siegmund Prillwitz*, pages 268–277. Signum, Hamburg.

Aylin C. Küntay and Dan I. Slobin. 1996. Listening to a turkish mother: Some puzzles for acquisition. In *Social Interaction, Social Context, and Language. Essays in the Honor of Susan Ervin-Tripp*, pages 265–286. Lawrence Erlbaum, Mahwah, NJ.

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

Francisco Lacerda. 2009. On the emergence of early linguistic functions: A biological and interactional perspective. In M. Lindgren M. Roll K. Alter, M. Horne and J. von Koss Torkildsen, editors, *Brain Talk: Discourse with and in the brain*, pages 207–230. Media-Tryck, Lund, Sweden.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Anat Ninio, Catherine E. Snow, Barbara A. Pan, and Pamela R. Rollins. 1994. Classifying communicative acts in children's interactions. *Journal of Communicative Disorders*, 27:157–187.

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

Robert Östling. 2013. Stagger: an open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

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

Heidi R. Waterfall. 2006. *A Little Change is a Good Thing: Feature Theory, Language Acquisition and Variation Sets*. Ph.D. thesis, Department of Linguistics, University of Chicago.

Mats Wirén, Kristina Nilsson Björkenstam, Gintare Grigonyte, and Elisabet Eir Cortes. 2016. Longitudinal studies of variation sets in child-directed speech. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, August 11, 2016, Association for Computational Linguistics*, pages 44–52, Berlin, Germany.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, pages 1556–1559, Genoa, Italy, May. ELRA.