Getting Creative: A Language Modeling Approach to Predicting Child Utterances in 12 Typologically Diverse Languages

Olivier Rüst, Marco Baroni, and Sabine Stoll

It has been proposed that in early development children rely on partially productive speech patterns and formulas, to infer increasingly more abstract grammatical representations (Arnon & Clark 2011; Lieven et al. 2009; Bannard & Matthews 2008; Goldberg 2006; Dąbrowska & Lieven 2005; Tomasello 2005; Lieven et al. 2003, 1997). Children have been shown to be sensitive to the probabilistic co-occurrence of speech units (Romberg & Saffran 2010; Saffran et al. 1999), which allows them to extract speech patterns from the language they hear. If, for example, two linguistic units often occur together, a child may analyze these as a coherent chunk. If this chunk is then followed by a more variable slot this gives rise to a partially productive schema: e.g., [*what's*][_]. In this case *what's* is invariant but can be followed by a limited set of words. Such small partially productive speech formulas are heavily featured in child and child-directed language (Moran et al. 2019, 2018; Lew-Williams et al. 2011; Stoll et al. 2009; Fernald & Hurtado 2006; Mintz 2003; Cameron-Faulkner et al. 2003; Redington et al. 1998). Here, we try to precisely map the course of child language from more formulaic to fully productive.

There is ample evidence that children combine such patterns, to create novel utterances. By joining formulas and expanding productivity in given "slots", an increasingly more creative language slowly emerges (Koch et al. 2020; Isbilen et al. 2020; McCauley & Christiansen 2019a, 2017; Chater et al. 2016; McCauley & Christiansen 2014; Bannard & Lieven 2012; McCauley & Christiansen 2011; Lieven et al. 2009; Bannard et al. 2009). So far, this line of research has focused on a limited set of speech formulas only and the relation to adult surrounding speech is not always clear. Moreover, most research is done on English and German and cannot be generalized to other languages due to a lack of cross-linguistic footing (with the notable exception of McCauley & Christiansen 2019a,b).

Here, we introduce a novel method to investigate this phenomenon in its general form. Our aim is to quantify language creativity as a whole. To achieve this, we rely on the idea of predictability of new utterances given previously used and

* Olivier Rüst, University of Zürich, ruest.olivier@uzh.ch. Marco Baroni, ICREA / Universitat Pompeu Fabra, mbaroni@gmail.com. Sabine Stoll, University of Zürich, sabine.stoll@uzh.ch.

 Proceedings of the 47th annual Boston University Conference on Language Development, ed. Paris Gappmayr and Jackson Kellogg, 672-685. Somerville, MA: Cascadilla Press.

heard ones. For this, we investigate patterns in child speech, adult speech and relate these in order to evaluate from where children receive their linguistic building blocks. We investigate longitudinal corpora of 12 typologically maximally diverse languages (Moran et al. 2016). Such a diverse sample of languages simulates the extreme variation of grammatical structures in the languages of the world and allows for language general conclusions.

We use a probabilistic language model to quantify the productivity of language as a whole. A language model is essentially a probability distribution of words, that allows to determine how probable a given sequence of words is. Our method is based on the intuition that such a probabilistic language model should be more successful at prediction the more formulaic speech is. In formulaic language words are more probable to appear in the same environment, which is comparatively easier to predict. If speech is fully productive, however, then words are much less likely to appear in the same vicinity and hence they are harder to predict. We use a neural language model based on the LSTM network architecture (Hochreiter & Schmidhuber 1997).

We evaluate child language, adult language and their interaction in three ways as the child grows up. We longitudinally predict (1) *child* utterances by *child* utterances, (2) *adult* utterances by *adult* utterances and (3) *child* utterances by *adult* utterances. This evaluates (1) the development of flexibility of child speech as they grow up, (2) how flexible surrounding adult speech is at the same time and (3) how "close" adult and child speech are to each other, i.e., how similar their structures and patterns are.

We have clear hypotheses concerning these predictions: (1) The predictability of *child* utterances by *child* utterances decreases, because children use gradually more complex speech formulas until they arrive at a fully creative language. (2) The predictability of *adult* utterances by *adult* utterances stays roughly the same or decreases slightly, due to adults already using a fully productive language that they may slightly adjust to children's needs. (3) The predictability of *child* utterances by *adult* utterances slowly decreases. This can be explained by the fact that young children use patterns they hear in the input but gradually adjust these to their own linguistic needs, as opposed to copying adult speech formulas.

In the following sections we show the results for these hypotheses and discuss their implications.

1. Predictability of Utterances

In order to capture as wide a variety of speech formulas as possible, we need a powerful method to recognize any probabilistic co-occurrence of words. Such co-occurrences can best describe speech formulas, no matter their exact shape. We rely on language models because these can learn such probabilistic co-occurrences of words. Take, as an instance, the example formula [*what's*] [_] [*doing*]. In this case, there is a very high probability (and is thus easily predictable) that *what* is followed by *'s*. The following slot is more variable, so that any element there has

a low transitional probability from *'s* or *doing*. Importantly, however, *doing* has a very high probability to appear after an intermediary element, such as for instance *she*, and is thus also very predictable.

Relying on the notion of predictability, then, allows us to quantify the extent to which a corpus of speech is largely formulaic. A language that contains a high degree of speech formulas is comparatively more predictable than one that is fully creative.

The exact choice of which type of language model to use is in so far irrelevant, as long as it is of good quality and maintained the same across measurements. Certain language models may outperform others on a specific dataset, but a longitudinal trend should be visible in all. That is, the model type may influence overall performance, but the longitudinal trend is maintained. In particular, we use the LSTM neural language model because it has been shown to capture more flexible contexts and outperforms classic count-based models (Graves 2012). At the same time, it is more computationally manageable than more recent Transformer-based models, which are most powerful in very high data regimes, which is not our case (Radford et al. 2019).

In a preliminary step, LSTM model hyperparameters (batch size, layers, sequence length, word embedding, learning rate, hidden dimensions, epochs) are set using a separate sample for each corpus. This sample consists of equal amounts of child speech and child-directed speech measured in words. This is done in order to ensure that the model hyperparameters do not favor the prediction of either child or child-directed speech. Subsequently, the model is applied longitudinally to the corpora as described in figure 1.

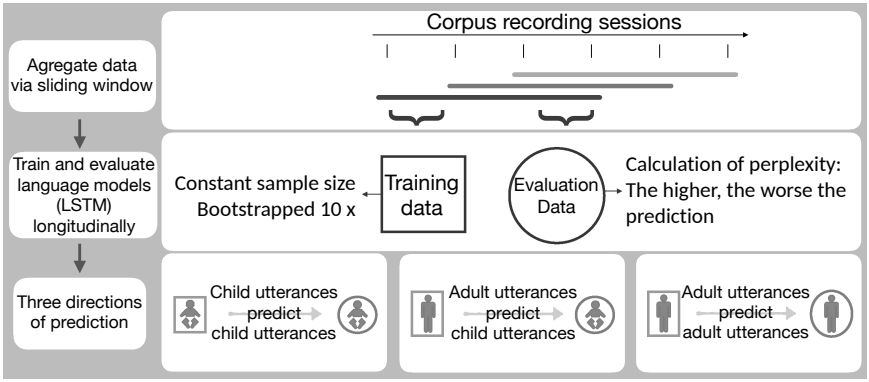

Figure 1. Pipeline to evaluate predictability of utterances.

First, a window of a certain length (dependent on corpus size, see below) is laid over the recording sessions of the corpora to aggregate sufficient data (these windows contain between 18,000 and 22,000 words, as these were the largest samples possible for the smaller corpora that still allow for longitudinal exploration).

This window is then moved along the time axis. For example, we determine a window of a length of 4 recording sessions. The first window then covers recording session 1 - 4. Then it is moved once so it covers session 2 - 5 and so on. The first half of the sessions within this window is used as the training set, the second as evaluation set. In order to make sure that amount of training or evaluation data does not influence the results, the number of words in both sets is kept constant over time via random downsampling to the highest possible value (i.e., the smallest sample found in any window of a given corpus). In order to still use all data, however, the process of training and evaluation is repeated (bootstrapped) 10 times, i.e., 10 models are trained and evaluated in each window. To evaluate the predictability of utterances of the models trained in this way, we use perplexity, the standard measure to quantify language models' performance. Perplexity measures lack of predictability, i.e., the higher the perplexity, the lower the predictability of utterances. In effect, in any given window, perplexity estimates are averaged across the 10 down-samplings of each window.

Moreover, there are different directions of prediction between child and adult speech: (1) *child* utterances predicted by *child* utterances, (2) *adult* utterances predicted by *adult* utterances and (3) *child* utterances predicted by *adult* utterances.

The results of these three directions of prediction are presented in the section after the description of our data.

2. Corpora

We use the ACQDIV database that contains a sample of typologically maximally diverse languages (Moran et al. 2016). These corpora are recorded longitudinally in order to capture the linguistic development of children. The languages of this database reflect the natural variety of languages that children have to face in the real world. Only such a diverse sample allows for language-general conclusions on acquisition. Information on the sub-corpora we use are given in the following table (Chintang: Stoll et al. 2015, Cree: Brittain 2015, English: Theakston et al. 2001, Indonesian: Gil & Tadmor 2007, Japanese: Nisisawa & Miyata 2009, 2010; Miyata & Nisisawa 2009, 2010; Miyata 2004a,b,c, 2012, Ku Waru: Rumsey & Yam 2019, Qaqet: Hellwig & Reetz 2014, Russian: Stoll & Meyer 2008, Sesotho: Demuth 2015, 1992, Tuatschin: Mazara et al. unpublished Turkish: Küntay et al. Unpublished, Yucatec: Pfeiler Unpublished).

Table 1.

Language	*Number of children*	*Age range (years)*	*Recording rhythm*	*Recording environment*	*Other speakers*
Chintang (Sino-Tibetan)	4	2 - 4	4h	outside, close to home	relatives, other children, passers-by
Cree (Algic)	1	2 - 3	30 - 40 mins every 2 - 3 weeks	indoors at home	mainly mother
English (Indo-European)	4	2-5	60 mins in every 3 week period	indoors at home	mainly mother
Indonesian (Austronesian)	3	1.5 - 3	40 - 60 mins every week	indoors at home	mainly mother
Japanese (Japonic)	7	1.5 - 5	indoors at home	70 min per week until 3, later every 1 or 2 months	mainly mother
Ku Waru (Trans–New Guinea)	1	2 - 3	160-65 minutes per month	indoors at home	father, mother, other children
Qaqet (Baining)	3	2 - 3	60 minutes per week	inside and outside home	parents, other children and adults

Language	*Number of children*	*Age range (years)*	*Recording rhythm*	*Recording environment*	*Other speakers*
Russian (Indo-European)	5	1- 7	60 mins every week	indoors at home	mother and relatives
Sesotho (Niger-Congo)	4	2 - 4	3 - 4 hours every month	home and vicinity	relatives, other children, passers-by
Tuatschin (Indo-European)	3	2 - 3	1h every week	indoors at home	mainly mother
Turkish (Turkic)	6	1 -3	1h every 2 weeks	indoors at home	mainly mother
Yucatec (Mayan)	3	1 - 3	30 - 90 mins every 2 weeks	home and vicinity	relatives

3. Results

Here we present the results for our three test cases: First, we evaluate how creative novel child utterances are, compared to utterances they have used before. Second, we evaluate the creativity of adult utterances, which serves as a baseline. Third, we evaluate how creative child utterances are in comparison to adult utterances, i.e., to what degree they rely on the structures that they hear from adults.

For all three directions of prediction we use the same model specifications. We use a Bayesian linear model to predict the language models' performance (measured via perplexity) with an interaction between child age and language, with random intercepts and slopes for speaker and language. We use the half-Cauchy distribution as a prior, as this gives more credibility to lower coefficients (Williams et al. 2018). Figure 2 shows an example regression line on perplexity over age in Chintang, predicting child utterances by child utterances.

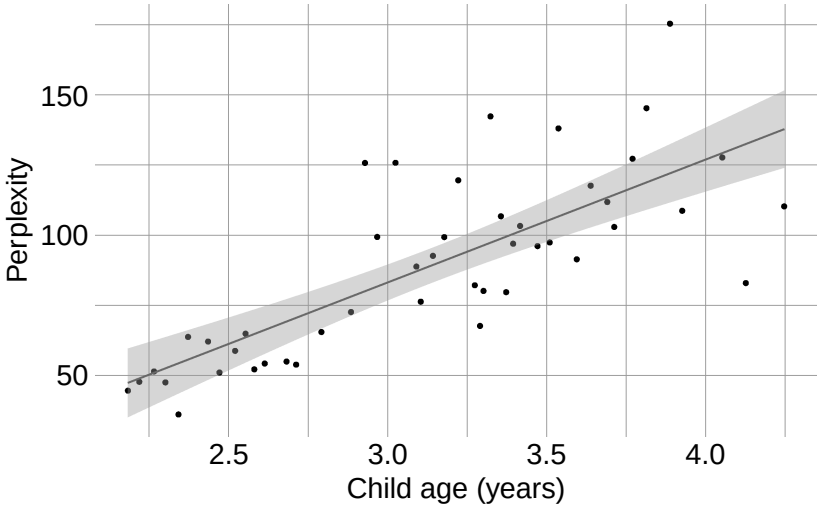

Figure 2. Example regression line of Perplexity over the age of the child. This illustrates the increase in model perplexity over time, predicting Chintang child utterances by child utterances.

Essentially, our model specifications account for such a regression line for each language separately. In figures 3, 4 and 5, the rows show the coefficients for the interaction between the specific language and age, i.e., the slope at which perplexity develops as the child grows older in each language. The point estimate is surrounded by the 95% credibility interval (in black) and the posterior distribution (in gray).

3.1. Child Utterances

First, the results of the model that predicts child utterances by child utterances are presented in Figure 3. There are three main observations: First, most languages show a statistically credible positive coefficient. This indicates that the language models' perplexity increases over time. Second, the smallest corpora (Tuatschin, Sesotho, Qaqet and Ku Waru) show the largest variance. This is expected as less data complicates the determination of an exact point estimate. Third, the rather large credibility interval for the smallest corpora (Qaqet and Ku Waru) contain 0. This indicates that there is a lot of uncertainty concerning the slope for the language model's perplexity and it cannot concluded that the slope is indeed positive.

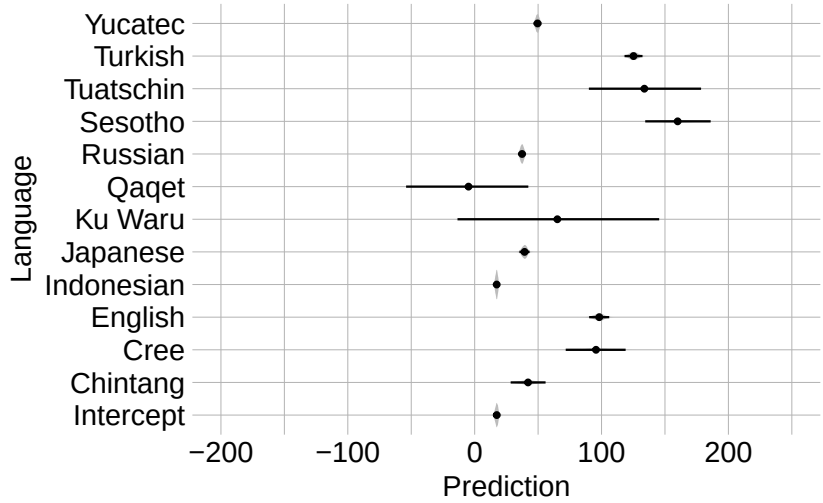

Figure 3. Predictability of *child* utterances by *child* utterances. The plot shows the coefficients for the interaction between child age and language, i.e., the slope at which the model's performance decreases over time. Around the point estimate the 95% credibility interval is indicated in black and the posterior distribution in gray.

Overall, however, most languages, excluding those for which the sample is very small, show a credible increase in model perplexity over time. This means that child language becomes less predictable based on what she has said before. More concretely, it indicates that the probabilities that describe the co-occurrence of words decreases over time. This is evidence that children use less formulaic and more creative language as they grow older.

3.2. Adult Utterances

The results of the model predicting adult utterances by adult utterances are presented in Figure 4. This type of prediction serves to illustrate the adults' behavior and is needed to interpret the prediction in the next section. There are 4 main observations: Generally, there are much lower - but still positive - coefficients in this type of prediction than in the previous model predicting child utterances by child utterances (except Qaqet, but the credibility intervals are rather large). This suggests that model perplexity also increases over time, but not as quickly as in the previous type of prediction. Second, the credibility intervals for the smaller corpora are, as expected, also large here. In combination with generally lower coefficients this results in the credibility intervals for Qaqet, Ku Waru and Chintang to contain 0. Third, the coefficients for Tuatschin and Cree are, surprisingly, negative. This indicates that model perplexity decreases over time. It is hardly credible

that adult language becomes less creative as a child grows older. We thus assume these coefficients reflect idiosyncrasies from the respective corpora deriving from their small size and sampling procedures.

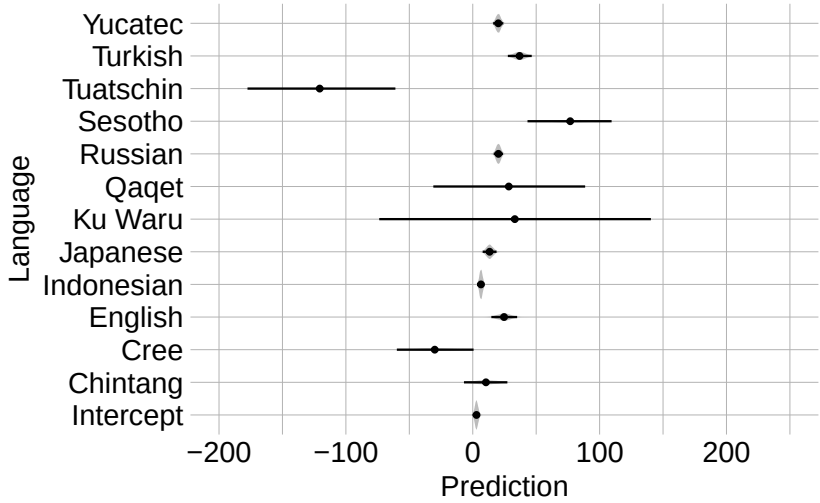

Figure 4. Predictability of *adult* utterances by *adult* utterances. The plot shows the coefficients for the interaction between child age and language, i.e., the slope at which the model's performance decreases over time. Around the point estimate the 95% credibility interval is indicated in black and the posterior distribution in gray.

For most languages, however, there are credible positive coefficients. This indicates an increase in creativity in adult language over time. This effect is, however, not as pronounced as in children. We understand this as evidence that adults accommodate to the linguistic needs of children. Concretely, with young children more formulaic language is used, which gradually becomes more creative as children understand more and use increasingly more creative language themselves.

3.3. Interaction of Child and Adult Utterances

Figure 5) reports the results of the model predicting child utterances by adult utterances. This type of prediction evaluates how good of a predictor adult speech is for child speech. In other words, it shows how similar adult and child speech formulas are. Essentially, perplexity in this instance is a measure of difference. A high similarity is indicated by low perplexity, whereas a higher perplexity indicates a greater divergence of speech formulas.

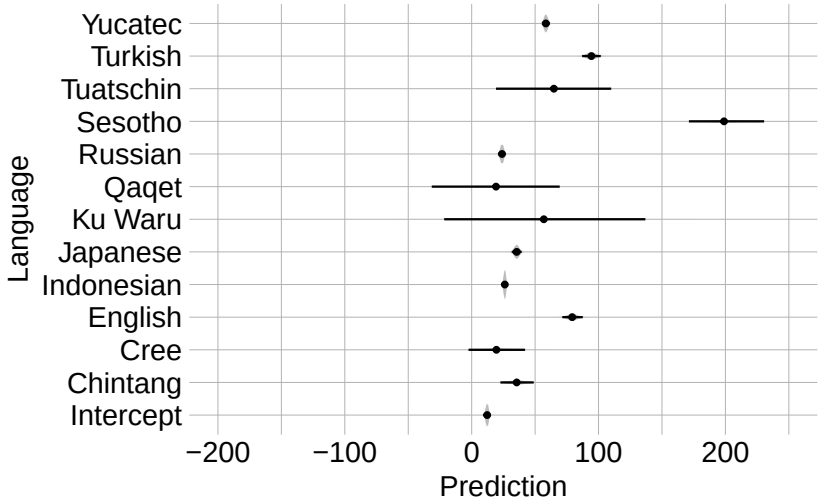

Figure 5. Predictability of *child* utterances by *adult* utterances. The plot shows the coefficients for the interaction between child age and language, i.e., the slope at which the model's performance decreases over time. Around the point estimate the 95% credibility interval is indicated in black and the posterior distribution in gray.

There are 3 main observations in this figure: First, the coefficients are mostly positive, indicating a continuous increase in perplexity over time. This indicates that adult speech becomes a continuously worse predictor for child language over time. This shows that children do not simply copy adult speech formulas (in which case we would expect a decrease in perplexity towards 0 because child and adult language become more similar, i.e., a negative coefficient). This may be explained by the fact that children's emergence of creativity is not only dependent on comprehending and copying structures from the input. It could indicate that an interplay of production as well as comprehension accounts best for the emergence of more productive formulas, where the reliance on adult speech diminishes slowly over time. Second, the credibility intervals for the small corpora are, as above, rather large. This results in the third main observation that the credibility intervals of Qaqet, Ku Waru and Cree to contain 0. If they are indeed 0, this would indicate that the degree of creativity of adult and child language remains constant (e.g., would increase their degree of creativity in parallel fashion).

Overall, however, children appear to develop their own set of speech formulas over time. A positive coefficient indicates that adult speech is a comparatively good predictor for early child speech. But, as the child grows older, the quality of prediction diminishes. We can tentatively conclude that children rely less on their input to create novel utterances the older they become. This is interesting as it

shows that children are able to extract grammatical information from the language they hear, and apply this knowledge to their own needs to interact with the world around them without copying the exact surface structure they encountered before. A more formal investigation is, however, necessary to evaluate the importance of comprehension and production for the emergence of creative formulas.

4. Discussion

Children have been shown to extensively use restricted speech formulas. We investigate how children go from these partially productive speech formulas to a fully productive language. In order to capture any possible speech formulas we present a novel method based on the predictability of utterances based on previously heard and used ones.

We build on the *Traceback method* (e.g., Lieven et al. 2009) and the *Chunk-Based learner* (e.g., McCauley & Christiansen 2017) that have previously been used to investigate speech formulas. In essence, the advantages of these are twofold: First, the *Traceback method* can capture non-sequential frame-and-slot patterns like [I want to][_][it] to reconstruct novel utterances. Second, the *Chunk-Based learner* learns sequential multi-word strings purely based on distributional information. This is done via backwards transitional probabilities, in order to predict novel utterances. We combine these advantages by using neural probabilistic language models and apply these longitudinally (cf. Koch et al. 2022; McCauley & Christiansen 2019b). These models can (1) learn any non-sequential patterns and (2) purely rely on distributional information (i.e., probability of co-occurrence of words).

In the first step, we show that child speech becomes continuously less predictable based on what she has said previously. The comparatively high probability of specific word co-occurrences in young children describe relatively stable speech formulas. By showing that these speech formulas are relaxed over time, we give evidence that child speech becomes less formulaic over time.

In the second step, we find that adults slowly relax the speech formulas that they adhere to. The opening of these formulas does, however, occur at a lower rate than the children's one. We suggest that this indicates that adults adjust to the linguistic needs of children. At first, they use more strict speech formulas to be able to convey meaning to the not yet linguistically competent children. These formulas are then relaxed as children grow to be more proficient speakers.

In the third step, we explore how divergent adult and child speech formulas are. The results indicate that young children rely much more on the input they receive from adults, i.e., their speech formulas are comparatively similar. Over time, however, these formulas gradually diverge, indicating that children use more productive language according to their own linguistic needs, without simply copying adult speech formulas. This may indicate that the reliance on adult input to create novel utterances diminishes over time.

To summarize, we show that (1) children gradually expand their speech formulas, (2) adults do likewise, in order to adjust their language to a child's needs, (3) that children rely less on the input for productive language the older they become.

Utilizing a sample of 12 typologically very diverse languages we can thus show that the incremental emergence of productivity forms a key building block of language acquisition. Emerging evidence suggests that children learn from very specific environments in adult language (e.g., child-directed speech (e.g., Matychuk 2005), variation sets (e.g., Moran et al. 2019), contingent speech (e.g., Elmlinger et al. 2022)). Our method of finding probabilistic speech formulas lends itself especially for the investigation of the origin of child speech formulas in such narrower linguistic environments.

References

Arnon, Inbal & Eve V Clark. 2011. Why brush your teeth is better than teeth–Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development* 7(2). 107–129.

Bannard, Colin & Elena Lieven. 2012. Formulaic language in L1 acquisition. *Annual Review of Applied Linguistics* 32. 3–16.

Bannard, Colin, Elena Lieven & Michael Tomasello. 2009. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences* 106(41). 17284–17289.

Bannard, Colin & Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological science* 19(3). 241–248.

Brittain, Julie. 2015. *Corpus of the Chisasibi Child Language Acquisition Study (CCLAS)*.

Cameron-Faulkner, Thea, Elena Lieven & Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive science* 27(6). 843–873.

Chater, Nick, Stewart M McCauley & Morten H Christiansen. 2016. Language as skill: Intertwining comprehension and production. *Journal of Memory and Language* 89. 244–254.

Dąbrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions.

Demuth, Katherine. 1992. The acquisition of Sesotho. In *The crosslinguistic study of language acquisition*, 557–638. Psychology Press.

Demuth, Katherine. 2015. *Demuth Sesotho Corpus*.

Elmlinger, Steven L, Michael H Goldstein & Marisa Casillas. 2022. Immature vocalizations simplify the speech of Tseltal Mayan and US caregivers. *Topics in Cognitive Science*.

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

Gil, David & Uri Tadmor. 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*.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Graves, Alex. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin: Springer.

Hellwig, Carmen Dawuda Henrike Frye, Birgit & Steffen Reetz. 2014. *The Qaqet Corpus at the Language Archive Cologne*.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780.

Isbilen, Erin S, Stewart M McCauley, Evan Kidd & Morten H Christiansen. 2020. Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science* 44(7). e12848.

Koch, Nikolas, Stefan Hartmann & Antje Endesfelder Quick. 2022. Traceback and Chunk-Based Learning: Comparing Usage-Based Computational Approaches to Child Code-Mixing. *Languages* 7(4). 271.

Koch, Nikolas, Antje Endesfelder Quick & Stefan Hartmann. 2020. Individual differences in discourse priming: A traceback approach. *Belgian Journal of Linguistics* 34(1). 186–198.

Küntay, Aylin Copty, Dilara Koçbaş & Süleyman Sabri Taşçı. Unpublished. Koç University Longitudinal Language Development Database on language acquisition of 8 children from 8 to 36 months of age.

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

Lieven, Elena, Heike Behrens, Jennifer Speares & Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of child language* 30(2). 333–370.

Lieven, Elena, Dorothé Salomo & Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis.

Lieven, Elena VM, Julian M Pine & Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of child language* 24(1). 187–219.

Matychuk, Paul. 2005. The role of child-directed speech in language acquisition: a case study. *Language sciences* 27(3). 301–379.

Mazara, Jekaterina, Jeraldine Walther & Sabine Stoll. unpublished. *Audiovisual corpus of Tuatschin*.

McCauley, Stewart M & Morten H Christiansen. 2011. Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In *Proceedings of the annual meeting of the cognitive science society*, vol. 33 33.

McCauley, Stewart M & Morten H Christiansen. 2014. Acquiring formulaic language: A computational model. *The mental lexicon* 9(3). 419–436.

McCauley, Stewart M & Morten H Christiansen. 2017. Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science* 9(3). 637–652.

McCauley, Stewart M & Morten H Christiansen. 2019a. Language learning as language use: A cross-linguistic model of child language development. *Psychological review* 126(1). 1.

McCauley, Stewart M & Morten H Christiansen. 2019b. Modeling Children's Early Linguistic Productivity Through the Automatic Discovery and Use of Lexically-based Frames. In *CogSci*, 782–788.

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

Miyata, Susanne. 2004a. Aki Corpus. *Pittsburgh, PA: TalkBank*.

Miyata, Susanne. 2004b. Ryo Corpus. *Pittsburgh, PA: TalkBank*.

Miyata, Susanne. 2004c. Tai Corpus. *Pittsburgh, PA: TalkBank*.
Miyata, Susanne. 2012. Japanese CHILDES: The 2012 CHILDES manual for Japanese .
Miyata, Susanne & Hiro Yuki Nisisawa. 2009. *MiiPro – Asato Corpus*. Pittsburgh, PA: TalkBank.
Miyata, Susanne & Hiro Yuki Nisisawa. 2010. MiiPro – Tomito Corpus. *Pittsburgh, PA: TalkBank*.
Moran, Steven, Damián E Blasi, Robert Schikowski, Aylin C Küntay, Barbara Pfeiler, Shanley Allen & Sabine Stoll. 2018. A universal cue for grammatical categories in the input to children: Frequent frames. *Cognition* 175. 131–140.
Moran, Steven, Nicholas A Lester, Heath Gordon, Aylin Küntay, Barbara Pfeiler, Shanley Allen & Sabine Stoll. 2019. Variation sets in maximally diverse languages. In *Proceedings of the 43rd annual Boston University Conference on Language Development*, 427–440.
Moran, Steven, Robert Schikowski, Danica Pajović, Cazim Hysi & Sabine Stoll. 2016. The ACQDIV database: Min (d) ing the ambient language.
Nisisawa, Hiro Yuki & Susanne Miyata. 2009. MiiPro – Nanami Corpus. *Pittsburgh, PA: TalkBank*.
Nisisawa, Hiro Yuki & Susanne Miyata. 2010. MiiPro – ArikaM Corpus. *Pittsburgh, PA: TalkBank*.
Pfeiler, Barbara. Unpublished. *Pfeiler Yucatec Child Language Corpus*.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8). 9.
Redington, Martin, Nick Chater & Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science* 22(4). 425–469.
Romberg, Alexa R & Jenny R Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(6). 906–914.
Rumsey, Andrew Noma Lauren Reed Naomi Peck Charlotte van Tongeren, Alan & Stephanie Yam. 2019. *ACQDIV portion of the Ku Waru Child Language Socialization Study (KWCLSS)*.
Saffran, Jenny R, Elizabeth K Johnson, Richard N Aslin & Elissa L Newport. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition* 70(1). 27–52.
Stoll, Sabine, Kirsten Abbot-Smith & Elena Lieven. 2009. Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech. *Cognitive Science* 33(1). 75–103.
Stoll, Sabine, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski & Balthasar Bickel. 2015. *Audiovisual corpus on the acquisition of Chintang by six children*.
Stoll, Sabine & Roland Meyer. 2008. Audio-visual longitudinal corpus on the acquisition of Russian by 5 children.
Theakston, Anna L, Elena VM Lieven, Julian M Pine & Caroline F Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language* 28(1). 127–152.
Tomasello, Michael. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.
Williams, Donald R, Philippe Rast & Paul-Christian Bürkner. 2018. Bayesian meta-analysis with weakly informative prior distributions.

Proceedings of the 47th annual Boston University Conference on Language Development

edited by Paris Gappmayr
and Jackson Kellogg

Cascadilla Press Somerville, MA 2023

ISSN 1080-692X
ISBN 978-1-57473-087-6 (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