L A N G U A G E D E V E L O P M E N T R E S E A R C H

An Open Science Journal

Volume 2 | Issue 1 | December 2022
ISSN 2771-7976

About the journal

Language Development Research: An Open-Science Journal was established in 2020 to meet the field's need for a peer- reviewed journal that is committed to fully open science: LDR charges no fees for readers or authors, and mandates full sharing of materials, data and analysis code. The intended audience is all researchers and professionals with an interest in language development and related fields: first language acquisition; typical and atypical language development; the development of spoken, signed or written languages; second language learning; bi- and multilingualism; artificial language learning; adult psycholinguistics; computational modeling; communication in nonhuman animals etc. The journal is managed by its editorial board and is not owned or published by any public or private company, registered charity or nonprofit organization.

Child Language Data Exchange System

Language Development Research is the official journal of the **TalkBank system**, comprising the CHILDES, PhonBank, HomeBank, FluencyBank, Multilingualism and Clinical banks, the CLAN software (used by hundreds of researchers worldwide to analyze children's spontaneous speech data), and the Info-CHILDES mailing list, the de-facto mailing list for the field of child language development with over 1,600 subscribers.

Diamond Open Access

Language Development Research is published using the Diamond Open Access model (also known as "Platinum" or "Universal" OA). The journal does not charge users for access (e.g., subscription or download fees) or authors for publication (e.g., article processing charges).

Clarifying revision made upon discovery of errata. Reissued on 20 December 2023.

Hosting

The **Carnegie Mellon University Library Publishing Service** (LPS) hosts the journal on a Janeway Publishing Platform with its manuscript management system (MMS) used for author submissions.

License

Peer Review and Submissions

All submissions are reviewed by a minimum of two peer reviewers, and one of our Action Editors, all well- established senior researchers, chosen to represent a wide range of theoretical and methodological expertise. Action Editors select peer reviewers based on their expertise and experience in publishing papers in the relevant topic area.

Submissions and Publication Cycle

We invite submissions that meet our criteria for rigour, without regard to the perceived novelty or importance of the findings. We publish general and special-topic articles ("Special Collections") on a rolling basis to ensure rapid, cost-free publication for authors.

Language Development Research is published once a year, in December, with each issue containing the articles produced over the previous 12 months. Individual articles are published online as soon as they are produced. For citation purposes, articles are identified by the year of first publication and digital object identifier (DOI).

Table of Contents

Volume 2, Issue 1, 31 December 2022

223

Parents' hyper-pitch and low vowel category variability in infant-directed speech are associated with 18-month-old toddlers' expressive vocabulary

Audun Rosslund, Julien Mayor, Gabriella Óturai, Natalia Kartushina

doi: 10.34758/2022.0547

268

Large-scale study of speech acts' development in early childhood

Mitja Nikolaus, Eliot Maes, Jeremy Auguste, Laurent Prévot, Abdellah Fourtassi

doi: 10.34842/2022.0532

306

Wishes before ifs: mapping "fake" past tense to counterfactuality in wishes and conditionals

Maxime Alexandra Tulling, Ailís Cournane

doi: 10.34842/2022.0559

COVID-19 first lockdown as a window into language acquisition: associations between caregiver-child activities and vocabulary gains

Natalia Kartushina
MultiLing, Department of Linguistics and Scandinavian Studies, University of Oslo, Norway

Nivedita Mani
Georg-Elias-Müller Institute for Psychology, University of Göttingen, Germany
Leibniz ScienceCampus Primate Cognition, Germany

Aslı Aktan-Erciyes
Department of Psychology, Kadir Has University, Turkey

Khadeejah Alaslani
Department of Linguistics, Purdue University, USA

Naomi J. Aldrich
Department of Psychology, Grand Valley State University, USA

Alaa Almohammadi
Haifa Alroqi
Department of European Languages and Literature, King Abdulaziz University, Saudi Arabia

Lucy M. Anderson
Brigham Young University, Provo, Utah, USA

Elena Andonova
Research Center for Cognitive Science, New Bulgarian University, Bulgaria

Suzanne Aussems
Department of Psychology, University of Warwick, United Kingdom

Mireille Babineau
Laboratoire de Sciences Cognitives et Psycholinguistique, École normale supérieure, PSL University, France
Department of Psychology, University of Toronto, Canada

Mihaela Barokova
Center for Autism Research Excellence, Boston University, USA

Christina Bergmann
Language Development Department, Max Planck Institute for Psycholinguistics, The Netherlands

Cara Cashon
Department of Psychological and Brain Sciences, University of Louisville, USA

Stephanie Custode
University of Miami, USA

Volume 2, Issue 1, 31 December 2022

Cielke Hendriks
Language Development Department, Max Planck Institute for Psycholinguistics, The Netherlands

Tzipi Horowitz-Kraus
Educational Neuroimaging Group, Faculty of Education in Science and Technology, Faculty of Biomedical Engineering, Technion, Israel

Marina Kalashnikova
Basque Center on Cognition, Brain, and Language, Spain
IKERBASQUE, Basque Foundation for Science, Spain

Junco Kanero
Faculty of Arts and Social Sciences, Sabancı University, Turkey

Christina Keller
Centre for Language and Communication Research, Cardiff University, UK

Grzegorz Krajewski
Faculty of Psychology, University of Warsaw, Poland

Catherine Laing
Centre for Language and Communication Research, Cardiff University, UK

Rebecca A. Lundwall
Brigham Young University, Provo, Utah, USA

Magdalena Łuniewska
Karolina Mieszkowska
Faculty of Psychology, University of Warsaw, Poland

Luis Muñoz
Department of Psychology, University of Oslo, Norway

Karli Nave
Faculty of Psychology, University of Warsaw, Poland

Nonah Olesen
Department of Psychological and Brain Sciences, University of Louisville, USA

Lynn Perry
University of Miami, USA

Caroline Rowland
Language Development Department, Max Planck Institute for Psycholinguistics, The Netherlands
Donders Institute for Brain, Cognition & Behaviour, Radboud University, the Netherlands

Daniela Santos Oliveira
University of Tennessee, USA
Jeanne Shinskey
Department of Psychology, Royal Holloway University of London, United Kingdom

Aleksander Veraksa
Lomonosov Moscow State University, Russia

Kolbie Vincent
Department of Psychological and Brain Sciences, University of Louisville, USA

Michal Zivan
Educational Neuroimaging Group, Faculty of Education in Science and Technology, Faculty of Biomedical Engineering, Technion, Israel

Julien Mayor
Department of Psychology, University of Oslo, Norway

Abstract: The COVID-19 pandemic, and the resulting closure of daycare centers worldwide, led to unprecedented changes in children's learning environments. This period of increased time at home with caregivers, with limited access to external sources (e.g., daycares) provides a unique opportunity to examine the associations between the caregiver-child activities and children's language development. The vocabularies of 1742 children aged 8-36 months across 13 countries and 12 languages were evaluated at the beginning and end of the first lockdown period in their respective countries (from March to September 2020). Children who had less passive screen exposure and whose caregivers read more to them showed larger gains in vocabulary development during lockdown, after controlling for SES and other caregiver-child activities. Children also gained more words than expected (based on normative data) during lockdown; either caregivers were more aware of their child's development, or vocabulary development benefited from intense caregiver-child interaction during lockdown, or both. We discuss these results in the context of the extraordinary circumstances of the COVID-19 pandemic and highlight limitations of the study.

Keywords: COVID-19 pandemic; vocabulary development; book reading; passive screen exposure; multi-country

Corresponding author: Natalia Kartushina, MultiLing, Department of Linguistics and Scandinavian Studies, Faculty of Humanities, University of Oslo, Niels Henrik Abels vei 36, 0313, Oslo, Norway. Email: natalia.kartushina@iln.uio.no

ORCID ID: https://orcid.org/0000-0003-4650-5832

Citation: Kartushina, N., Mani, N., Aktan-Erciyes, A., Alaslani, K., Aldrich, N. J., Almohammadi, A., ... & Mayor. J. (2022). COVID-19 first lockdown as a window into language acquisition: associations between caregiver-child activities and vocabulary gains. *Language Development Research*, 2(1), 1—36. https://doi.org/10.34842/abym-xv34

Introduction

Language is acquired knowledge – children need experience with language to learn it. Differences in the quality and quantity of children's language experience may, therefore, influence language learning outcomes. Indeed, the *quantity* of children's language input is positively associated with their vocabulary size and development (in Western, industrialized societies, see Hart & Risley, 1995; Weisleder & Fernald, 2013; but see Casillas et al., 2020 for work on non-Western societies). The *quality* of children's language experience is similarly associated with language development with findings suggesting that the diversity, sophistication and responsiveness of input predict later vocabulary growth (Anderson et al., 2021; Cartmill et al., 2013; Pan et al., 2005).

Much of the work examining factors associated with variability in early language development (Frank et al., 2021) has focused on caregivers' reports of their daily interactions with their children. Such reports do not include input that children routinely receive from other sources (e.g., daycare centers, screen exposure), making it difficult to quantify all of the linguistic input available to children. In early 2020, the COVID-19 pandemic led many countries to implement strict lockdowns such that families had little or no social contact with others outside their household. Schools and daycare centers were shut down in over 160 countries (COVID-19 Educational Disruption and Response. *UNESCO*). Many caregivers worked from home, providing them with a better overview of their child's development and the activities their children were engaged in. Such periods of extended contact between caregivers and children have previously been referred to as "faucet" moments (Entwisle et al., 2001), when shared aspects of the child's environment, e.g., schools and daycare centers, are removed, such that differences in the home environment are particularly weighted in development. The current study capitalized on this "faucet" moment during the first COVID-19 lockdown to examine whether the activities that caregivers and children engaged in correlated with children's vocabulary development during this period.[1]

To achieve these goals, we evaluated, first, the amount of time children spent during lockdown on the following activities[2] (together with a caregiver or alone): shared book

[1] Throughout this manuscript, we will refer to lockdown as the time from March to September 2020 during which daycare centers were closed – and not in the sense of a strict curfew.

[2] To our knowledge, no questionnaire assessing parental activities has been validated across the populations examined in the current study, i.e., 13 countries with children learning 12 different languages. This required us to develop a questionnaire on the activities that caregivers undertook with their children during the COVID19-related lockdown. We acknowledge, however, that this questionnaire has not been validated across the populations tested. It is noteworthy that due to the extra-ordinary time constraints on data collection (the questionnaires needed to be approved by ethics board before launching the study, and sent out as soon as lockdown ended), and, given that children were

reading, structured child-caregiver games (referred to as structured parent-child interaction in the preregistration), free play with their caregiver, singing, speaking, outdoor activities, watching TV, baby shows or cartoons (henceforth, referred to as passive screen exposure), playing digital baby games (henceforth, active screen exposure involving interaction with a device), and playing freely without adults. Then, we assessed whether the time spent on these activities correlated with vocabulary development during lockdown, as indexed by the difference in the child's vocabulary size (in percentile, compared to norms, and in raw scores, where norms were not available) at the beginning and end of the lockdown period. To measure children's vocabulary sizes, we used Communicative Development Inventories (CDI; Fenson et al. 2007) – vocabulary checklists, where parents check words that their child understands or understands and produces. We focused on these activities given prior research finding positive associations between vocabulary development and shared book reading (Shahaeian et al., 2018; Wasik et al., 2016), speaking (Weisleder & Fernald, 2013; Rowe, 2018), singing (Williams et al., 2015), and playing (Hirsh-Pasek et al., 2009); and negative associations between screen exposure (van den Heuvel et al., 2019; Zimmerman et al., 2007) and vocabulary development.

In addition, we also measured caregiver's education (as a proxy for SES) to account for its potential associations with vocabulary development. Previous research suggests that children from higher-SES homes have larger vocabularies than those from lower-SES homes (Pace et al., 2017, Rowe, 2018). SES also moderates the relationship between caregiver-child activities and vocabulary development (Shahaeian et al., 2018, but see Malin et al., 2014). We chose maternal education as a proxy for SES because caregiver education is an important foundation for different developmental outcomes (Davis-Kean et al., 2020). We, therefore, statistically controlled for maternal education attainment in examining the association between caregiver-child activities and vocabulary development during lockdown. In addition, we examined the correlation between maternal education and the activities that caregivers engaged in.

We predicted (see https://osf.io/r85fw) that children whose caregivers engaged more in activities known to promote language development would have larger gains in receptive and productive vocabulary by the end of lockdown. In particular, we predicted that the frequency of shared book reading would capture more of the variability in vocabulary development than the frequency of other activities we examined (Montag et al., 2018), and that increased passive screen exposure would be related to smaller gains in vocabulary development (Zimmerman et al., 2007). Furthermore, we predicted that children whose caregivers engaged in more interactive shared book reading (e.g., asking questions, pointing to things) and structured caregiver-child games (Hirsh-Pasek et al., 2009) would show larger gains in vocabulary (Flack et al.,

already in lockdown when the study started, the questionnaires could not be normed - these were not typical circumstances.

2018). We predicted that children of caregivers with lower maternal education would have (a) smaller gains in both receptive and expressive vocabulary size over lockdown than children of caregivers with higher educational attainment, and (b) smaller vocabulary size at the start of lockdown (Pace et al., 2017; Rowe, 2018). However, we also predicted that the relationship between maternal education and vocabulary development would be better explained by the activities that caregivers engaged in with their children: while there may be differences in the activities that caregivers differing in educational attainment engage in with their children (Entwisle et al., 2001; Pace et al., 2017), the duration and the frequency of such activities should be associated with vocabulary gains during lockdown, above and beyond educational attainment. Finally, we also predicted that infants who attended kindergarten before the lock-down period might experience bigger changes in the quantity and quality of parent-child interactions (before vs. during lockdown) as compared to those who did not, which would translate into bigger changes in vocabulary size during lockdown for the former.

Methods

Participants

In total, 5494 caregivers - from 15 countries and 23 labs - filled in the Time 1 (T1) questionnaire at the beginning of lockdown in their country/region (see Supplementary Material 1 for additional sample details) and 2830 caregivers - from 14 countries and 21 labs - filled in the Time 2 (T2) questionnaire at the end of lockdown (see Procedure for details). Among the 2830 caregivers who filled in T2 questionnaires, data regarding 798 children were excluded from the analysis for either not meeting the following inclusion criteria: (a) monolingual children, defined as having a minimum of 90% exposure to their native language, according to caregiver reports, (b) full term babies, defined as born at 37 weeks of gestation or later, (c) no diagnosed developmental disorder, and (d) no hearing/vision impairment; or when we were unable to match participant ID and/or date of birth across both questionnaires. Note that data gathered for bilingual and multilingual children excluded from the present analysis will be analyzed in a separate contribution. In addition, we excluded 16 children who were outside the normative age range of country-specific CDIs (Fenson et al. 2007). Finally, upon careful analysis of the raw data, we excluded 79 children (2.5% of production and 4.4% of comprehension data), whose gains or losses per day in raw CDI comprehension or production scores fell outside of the distribution and were theoretically or practically uninterpretable for a typically developing child (see Analyses.Rmd code on https://osf.io/ty9mn/), likely due to parental inattentiveness or lack of involvement in the project (cf 7-13% exclusion of unreliable caregivers in de Mayo et al. (2021) for suspiciously brief CDI completion times).

Upon application of the inclusion criteria, our final sample comprised 1742 child participants[3] (886 girls and 856 boys; *M* age = 627 days at T1, range = 244-1089) from 18 labs and 13 countries that contributed to both T1 and T2 data. The SES varied between 1 (primary school, 0.57% of the data) and 6 (doctoral degree, 9.7% of the data), with the median education level of 4 (*SD* = 0.9), where 4 is Bachelor degree (27.78% of the data); these data suggest that mothers in this sample had, overall, high education levels, with the largest proportion of mothers having a MA degree (51.5%) and only 2.7% and 6.49% of the mothers having a high school and some college degree, respectively; although there were notable differences across countries (for details, see Analyses_2.html on https://osf.io/ty9mn/). Yet, note that, for the countries for which data on maternal educational attainment were available in wordbank.stanford.edu (Frank et al., 2017), the proportion of mothers with lower education levels (1 and 2 on the maternal education scale) was comparable to that reported in the normative data (see Supplementary Material 3), suggesting that the proportion of mothers with lower educational attainment in our sample was not lower than what can be found in the country-specific normative data, in general. An additional 290 children from Switzerland (for whom the exact age was missing) were included in the analyses of the relationship between SES and activities reported on https://osf.io/ty9mn/ (total *n* = 2033). Information about labs and child participants is included in Table 1.

Materials

T1 Questionnaire

The questionnaire launched at the beginning of lockdown included basic demographic questions about the children (sex, date of birth, estimated proportion of language exposure to each language heard in their daily life, preterm-versus-full-term status, history of ear infections, known hearing or visual impairments, and known developmental concerns), their caregivers (sex, level of education, and native language(s)) and siblings, if any (sex and date of birth). Maternal education (proxy for SES) was measured on a scale from 1 to 6, with 1 – primary school, 2 – high school, 3 – some college/university, 4 – Bachelor degree, 5 – Master degree, and 6 – doctoral degree (see https://osf.io/ty9mn/ for the distribution of maternal education in each country).

We measured children's receptive (for children between 8 and 18 months of age) and expressive (from 8 to 36 months of age) vocabularies at the onset of lockdown using age-appropriate CDIs and their adaptations for the relevant language (or regional variant). Variants included short-CDIs (Mayor & Mani, 2019 – for German) and web-CDIs (de Mayo et al., 2021— for American English, Hebrew, Dutch). CDIs ranged from 303

[3] Note that given that all questions had an option "prefer not to answer", some participants, in the final sample, had no data for some activities or SES.

to 897 words (25 items for the short-CDIs in German). A subset of laboratories collected additional data (not analyzed here) for use in planned follow-up projects.

Table 1. ***Description of the final sample of children (number, mean age in months and sd) included in the analyses of gains in production and comprehension (in percentile and raw CDI score).***

			Production (raw CDI score)		Comprehension (raw CDI score)		Production (percentiles)		Comprehension (percentiles)	
Labid	Country	Language	Age	n	Age	n	Age	n	Age	n
babyling	Norway	Norwegian	21 (6.9)	173	13.1 (2.7)	58	21 (6.9)	173	13 (2.7)	58
bcbl	Spain	Basque	17 (6.5)	18	12.5 (0.9)	10	NA	NA	NA	NA
bcbl	Spain	Spanish	15 (6.6)	19	9.8 (1.7)	10	NA	NA	NA	NA
brc-nijmegen	The Netherlands	Dutch	17 (6.8)	20	12.2 (3.7)	11	NA	NA	NA	NA
brookes	UK	English	19 (7.2)	292	12.6 (2.5)	143	15 (1.1)	83	15 (1.1)	81
clcu	UK	English	20 (7.6)	40	13.1 (3.6)	17	16 (1.6)	10	16 (1.5)	9
cogdevlabbyu	USA	English	12 (3)	39	12.1 (3.0)	38	12 (2.9)	36	12 (2.9)	35
dsc	USA	English	21 (6.6)	5	14.7 (1.3)	2	23 (6.6)	4	14	1
goe	Germany	German	21 (1.6)	37	NA	NA	21 (1.5)	36	NA	NA
HaifaUniv	Israel	Hebrew	21 (5.5)	61	13.5 (2.7)	12	15 (1.4)	11	15 (1.1)	9
ilpll	USA	English	21 (9.0)	49	11.2 (1.9)	16	16 (6.2)	32	11 (1.5)	15
kau-cll	Saudi Arabia	Arabic	22 (6.3)	90	11.3 (1.9)	10	NA	NA	NA	NA
ldl	Canada	English	22 (8.4)	17	12 (3.3)	5	20 (5.8)	12	13 (3.1)	4
mltlab	Turkey	Turkish	24 (6.2)	40	12.8 (2.3)	4	24 (5.5)	36	12 (1.7)	3
msu	Russia	Russian	22 (5.3)	17	15.9 (2.5)	4	23 (5.5)	14	14 (1.8)	2
multilada	Poland	Polish	21 (6.8)	223	13.6 (2.6)	77	21 (6.8)	209	13 (2.4)	69
paris_team	France	French	22 (6.8)	466	12.9 (1.9)	113	NA	NA	NA	NA
rhul_baby_lab	UK	English	15 (1.9)	25	14.4 (1.8)	22	15 (1.1)	23	15 (1.2)	21
technion_il	Israel	Hebrew	22 (7.1)	111	14 (2.5)	33	16 (1.8)	30	15 (1.7)	23
		Total		1742		585		709		330

Note. NA - not available, indicates when CDI norms were not available for a given language and/or CDI instrument. In the Brookes sample, 7 participants in the percentile analysis and 15 in the analysis of raw CDI were exposed to limited daycare during lockdown (means of 1.4 and 1.5 days a week, respectively).

T2 questionnaire

To assess activities that caregivers and their children engaged in during lockdown, a custom-made questionnaire was created and then collaboratively expanded and refined until the launch of the project. Questions evaluated the time spent on the following activities during lockdown: shared book reading, structured child-caregiver games, free play with the child, singing with the child, one-to-one speaking with the child, time spent outdoors, passive screen exposure (watching baby TV, cartoons, shows, with no interaction with a digital device), playing baby games on a digital device, time spent playing without an adult – all on a 10-point scale ranging from "did not do this activity at all" to "more than 4 hours most days." If parents/caregivers indicated that they read to their child at least 15 minutes per day, then they were asked eight yes/no questions (receiving each 1 point for a "yes" answer) on the quality of reading interactions (Han & Neuharth-Pritchett, 2015). The questionnaire also asked about the amount of time caregivers spent working from home and included CDI data to measure vocabulary development over the lockdown period. A subset of laboratories collected additional data (not analyzed here) for use in planned follow-up projects.

Procedure

On March 12, 2020, the Norwegian government enforced a national lockdown and, among other measures, closed daycare centers. On March 18, the local study on the impact of lockdown on language acquisition among 8- to 36-month-old children in Norway was preregistered and data collection started on March 20. The same day, a call for participation for international partners was issued via various mailing lists, which resulted in the present collaboration, including 23 labs in 15 countries. Each lab was asked to launch the T1 questionnaire as soon as possible upon daycare centers' closure and to launch T2 as close as possible to children starting regular daycare again, or if significant changes took place in local policies that would affect social isolation. Data collection started on March 20, 2020 (Norway) and finished on September 29, 2020 (USA), with a mean time interval between T1 and T2 of 41 days. We welcomed participation from all labs that were able to obtain ethical approval in time to launch the T1 questionnaire close to the daycare centers' closure. No minimum participant numbers were required to join the project.

The whole study was conducted online. We used a variety of means to recruit participants (e.g., social media, lab databases, social platforms, etc.), which allowed us to reach out to larger demographic populations, as compared to those typically tested in the lab (de Mayo et al., 2021). Data collection took part during the first COVID-19 lockdown. The announcement invited parents of 8-36-month-old infants to take part in a research project and included a link to the T1 questionnaire (see Materials), where caregivers were also asked to generate a unique participant identifier and provide a

valid email address, to be used when sending them the T2 questionnaire. Participant compensation varied across labs from no compensation to a small toy, a book or a voucher or a lottery ticket to win gift cards. The research project was approved by the Norwegian Center for Research Data REF536895 and by the ethics committee of the Department of Psychology at the University of Oslo. Collaborating labs obtained ethical approval from their institutions. Central data analyses used exclusively anonymized data.

Transparency statement

Prior to data collection, and prior to the call for an international collaboration, we preregistered our study for the Norwegian sample (https://osf.io/4mhjw). To accommodate for multi-site analyses, and to include modifications made to the questionnaires in the days following the initial preregistration, a multi-site preregistration was made prior to data inspection, visualization and processing (https://osf.io/r85fw). All materials, anonymized data, and analysis codes are available on the project's OSF (https://osf.io/ty9mn/).

Results

Data Processing

Computation of Vocabulary Gains in Percentiles

Our dependent variables were the total number of words that caregivers reported their child understood (between 8 and 18 months of age) and produced (between 8 and 36 months of age). The total number of words on CDIs was transformed into daily percentiles separately for each language using available norming data from wordbank.stanford.edu (Frank et al., 2017), provided that the dataset was dense enough, with a minimum of 50 data points per age (in months), or, for Hebrew, Polish and British English (UK-CDI), via direct contact with the authors who collected the norming data. Monthly percentiles from the norming data were linearly interpolated to establish daily percentiles (i.e., daily norms), then used to compute children's vocabulary size in daily percentiles (cf. https://osf.io/ty9mn/). We were able to derive daily percentiles for 14 labs in 9 countries (cf. Table 1) and computed gains in percentiles (T2-T1) for both comprehension (n = 330) and production (n = 709).

Computation of Normalized Gains in Raw CDI Scores

For 6 CDI instruments from 6 countries, data was either not available on WordBank (Saudi Arabia, the Netherlands, extended OxfordCDI) or the data available on Wordbank was too sparse to ensure reliable computation of percentiles (France, Spain, Is-

rael CDI - WS), despite children meeting the criteria for inclusion in the study. Therefore, these data were only entered into the analyses of raw CDI scores (along with the data from children that entered the percentile analyses).

Given (1) wide variation in the CDI size across languages (from 303 to 897) and (2) that potential gains were constrained by CDI scores at T1 (e.g., a toddler knowing all of the words on the CDI at T1 cannot learn more words on the CDI at T2), we computed a normalized measure of gains for each child that situated her with respect to the average gains from all countries given the same relative number of words known on her respective CDIs at T1 (see Analyses.Rmd on https://osf.io/ty9mn/). To this end, first, we divided the CDI score at T1 by the total number of items on the CDI, thus producing a vocabulary proportion score at T1, that varied between 0 and 1. Second, we fitted a polynomial regression to the T1 proportion score, separately for each tool (CDI Words and Gestures and CDI Words and Sentences) and modality, using the *loess* function and then used *predict* on the model outcomes to compute the average expected gains associated with T1 proportion scores. Then, we subtracted average expected gains associated with the T1 proportion scores from actual gains, resulting in average normalized gains of zero, for all T1 proportion scores (see Supplementary Material 2 for the visualization of non-normalized and normalized gains in vocabulary size). In other words, this procedure allowed us to identify individual deviations from expected gains (controlling for the CDI size and the CDI raw score at T1), and to correlate such deviations from expected gains with activities during social isolation. This normalization procedure for gains in raw CDI scores was conducted separately for each CDI tool and modality, for the entire sample comprising 18 labs from 13 countries in: comprehension (n = 585, 8-18-month-old children) and production (n = 1742, 8-36-month-old children).

Statistical Analyses

Correlations between SES and Activities

Pearson correlations (n = 709, dataset for the analyses of percentile gains in production) between SES and activities are reported in Table 2. Correlation matrix for a larger data set with n = 2033 children (that includes Switzerland and the labs for which norming data for the vocabulary score were not available) is available on the OSF page of the project https://osf.io/ty9mn/. As predicted, maternal education correlated positively with the time spent on shared book reading and negatively with children's passive screen exposure. Moderate correlations (>.30) included: a positive correlation between the time spent on shared book reading and on structured child-caregiver games, and between the time spent on passive screen exposure and playing baby games on a digital device. All other correlations were weaker (<.30). We hypothesized that the relationship between screen exposure and SES might be influenced by parents' availability, indexed by the number of hours they worked from home. A separate

linear model, however, revealed that this interaction was not significant (β = 0.0174, SE = 0.028, t = 0.62, p = 0.534).

Maternal education, activities and gains in production

First, a mixed-effect regression analysis on percentile gains in production, was conducted in R (R Core Team, 2020) for children between 8- and 36-months-of-age (see Table 3) using *lmer* (Bates et al., 2015:4) and *summ* (Long, 2020) to obtain the summary of the model. Fixed factors were time spent on activities that caregivers engaged in with their child during lockdown (mean-centered), maternal education (mean-centered), child's sex, and age (mean-centered in days, at T1), time gap between T1 and T2 in days (mean-centered), and child's daycare attendance before T1 (yes/no). Descriptive statistics for the activities and other variables used in the model can be found in the Supplementary Material 4. Random effects included a maternal education by country slope, hence, country was included as a random factor.[4]

Next, the same analysis was conducted on the second dependent variable, i.e., normalized raw gains in production. The results of the two models are summarized in Table 3. Note that the intercept and the effect of time gap between T1 and T2 need to be interpreted differently across the percentile and raw gains models. The intercept in the percentile model examines whether children (at the reference level of mean-centered age) gained more words than expected during lockdown (given normative data), since we expect children to stay in the same percentile across development. The intercept in the raw gains model is not meaningful as gains were normalized for each instrument. Time gap in the percentile model examines whether children's percentile scores improved linearly with the duration of lockdown, i.e., that they showed greater improvement in their percentile scores, the longer lockdown lasted. Time gap in the raw gains model trivially examines whether children learned more words the longer lockdown lasted.

[4] In order to address a potential issue of cryptic multiple testing raised by one of the reviewers, we performed, as recommended in Forstmeier & Schielzeth (2011), a full-null model comparison for both dependent variables (gains in percentiles and in normalized raw CDI scores), where the full model contained all the factors included in the main model and the null model excluded the activities examined in the paper. The results of the full/null comparison revealed a significant difference between the two models in gains in percentiles (x^2 = 17.6, df = 9, p = .04) and a marginal difference in gains in normalized raw CDI scores (x^2 = 16.2, df = 9, p = .063), suggesting that activities significantly improved the fit of the null model.

Table 2. ***Means, standard deviations, and correlations with confidence intervals between SES and activities.***

Variable	*M*	*SD*	Maternal education	Book reading	Caregiver works @home	Outdoor activities	Free play w.child	Singing	Speaking	Screen exposure	Digital games	Structured games
Maternal education	4.50	0.89										
Book reading	4.06	1.58	**.15**** **[.08, .22]**									
Parent works @home	3.54	3.51	.03 [-.04, .11]	-.03 [-.11, .04]								
Outdoor activities	4.47	2.67	.02 [-.05, .10]	-.00 [-.08, .07]	.03 [-.05, .10]							
Free play w. child	5.83	1.91	.07 [-.01, .14]	**.24**** **[.17, .31]**	.02 [-.06, .09]	**.16**** **[.09, .24]**						
Singing	3.72	1.74	-.03 [-.10, .05]	**.14*** **[.06, .21]**	.03 [-.05, .10]	.11 [.04, .19]	**.21**** **[.14, .28]**					
Speaking	5.94	2.13	-.03 [-.10, .05]	**.20**** **[.13, .27]**	-.04 [-.12, .03]	.04 [-.03, .12]	**.29**** **[.22, .35]**	**.28**** **[.21, .34]**				
Screen exposure	3.24	2.36	**-.16**** **[-.23, -.08]**	**-.12*** **[-.20, -.05]**	.06 [-.02, .13]	**.14**** **[.07, .21]**	-.01 [-.09, .06]	.03 [-.04, .11]	.03 [-.04, .11]			
Digital games	0.52	1.26	-.10 [-.17, -.02]	-.08 [-.16, -.01]	.05 [-.03, .12]	.06 [-.01, .13]	-.03 [-.10, .04]	.04 [-.03, .12]	.01 [-.07, .08]	**.33**** **[.26, .39]**		
Structured games	2.48	1.91	.04 [-.03, .11]	**.41**** **[.35, .47]**	-.07 [-.14, .01]	.04 [-.03, .11]	**.18**** **[.10, .25]**	**.17**** **[.10, .24]**	**.18**** **[.11, .25]**	.11 [.03, .18]	.06 [-.01, .14]	
Free play no adults	5.16	1.90	-.10 [-.17, -.03]	**-.16**** **[-.23, -.08]**	-.00 [-.08, .07]	.09 [.02, .17]	-.00 [-.08, .07]	-.00 [-.08, .07]	.06 [-.01, .14]	**.23**** **[.16, .30]**	**.14**** **[.06, .21]**	.01 [-.07, .08]

Note. * indicates $p < .05$. ** indicates $p < .01$. The Holm method was used to correct for multiple comparisons and adjust p-values.

Table 3. ***Fixed effects from the mixed-effect regression on the gains in production (left: percentiles with n = 709, full cases n = 685; right: raw scores with n = 1742). p-values below .05 are marked in bold.***

	Gains in percentiles					Normalized gains in raw CDI scores				
	Est.	SE	t	df	p	Est.	SE	t	df	p
(Intercept)	**3.32**	**1.10**	**3.01**	**685**	**.00**	4.01	3.87	1.04	13.45	.32
SES	-1.08	.68	-1.58	685	.11	-.88	2.22	-.40	8.12	.70
Book reading	.16	.43	0.38	685	.71	**1.71**	**.74**	**2.32**	**1528.0**	**.02**
Structured caregiver-child games	-.06	.37	-0.18	685	.85	.42	.57	.73	1601.0	.47
Passive screen exposure	**-.86**	**.29**	**-2.97**	**685**	**.00**	**-1.14**	**.50**	**-2.27**	**1377.8**	**.02**
Outdoor activities	-.09	.23	-0.40	685	.69	.17	.40	.43	1453.4	.67
Digital games	**1.08**	**.48**	**2.24**	**685**	**.03**	.15	.81	.19	1613.5	.85
Free play w. child	.29	.33	0.89	685	.37	-.42	.55	-.75	1616.7	.45
Singing	-.57	.35	-1.63	685	.10	.32	.63	.50	1603.5	.61
Speaking	.39	.29	1.34	685	.18	.17	.51	.34	1576.0	.74
Free play no adult	-.05	.32	-0.16	685	.88	.10	.51	.19	1618.3	.85
Time gap	-.02	.03	-0.65	685	.52	**.55**	**.07**	**8.14**	**150.4**	**.00**
Daycare before (yes)	1.18	1.22	.97	685	.33	1.05	2.51	.42	1201.8	.68
Gender (m)	.17	1.16	.15	685	.88	-1.37	1.93	-.71	1613.2	.48
Age (T1)	.00	.00	.52	685	.61	-.00	.01	-.33	1587.6	.74

Note. all numeric predictors were mean-centered in the analyses; p-values were calculated using Satterthwaite d.f.

In both analyses, the time spent on passive screen exposure negatively correlated with gains in productive vocabulary. As seen in Figure 1, children with no exposure to screens were reported to have the largest gains relative to the normative (age-matched) data from the CDI measures. Yet, it is noteworthy that regardless of the time spent on screen use, reported gains in production always exceeded or met expectations (a gain of zero is equivalent to what would be expected in the normative data).[5]

We also note that the intercept in the percentile model is significantly above zero, i.e., analyses of caregiver reports suggested that children (at the reference level of mean-centered age) gained more words in their productive vocabularies during lockdown, i.e., daycare closure, when compared to the normative data. A Wilcoxon signed-rank test with continuity correction found no evidence for a difference in children's reported vocabularies relative to normative data at the start of lockdown, at T1 (p = .5, Q1 = 23, median = 50, Q3 = 74), but larger reported vocabularies relative to normative data by the end of lockdown, at T2 (p = .005, Q1 = 28, median = 56, Q3 = 80). As indicated by the significant intercept, a one sample t-test on percentile gains between T1 and T2 revealed that, according to caregiver reports, children gained an average of 4 percentiles by the end of lockdown at T2 (95%CI = [2.7:5.0]; $t(684) = 7.0$, $p < .001$, $d = 0.26$).

The effect of time gap on the normalized gains in production suggests that caregivers reported that the longer the time gap between T1 and T2 was, the more words their children learned. In contrast, we found no evidence that percentile gains in vocabulary size accumulated over lockdown, i.e., that children showed greater vocabulary gains (relative to normative data), the longer lockdown lasted. The effects of digital media games on gains in percentiles, and of shared book reading on normalized raw gains did not replicate across analyses and will not be discussed further. Note also that a positive effect of digital media games on gains in percentiles should be interpreted with caution as 79% of children did not play digital games at all. There were no significant associations between gains in production and children's gender or age.

[5] As preregistered, we re-analyzed the data when >95% and <5% percentiles were excluded to check whether the model outcomes were impacted by these extreme values; the significant intercept and main negative effect of passive screen exposure remained significant (see details on OSF).

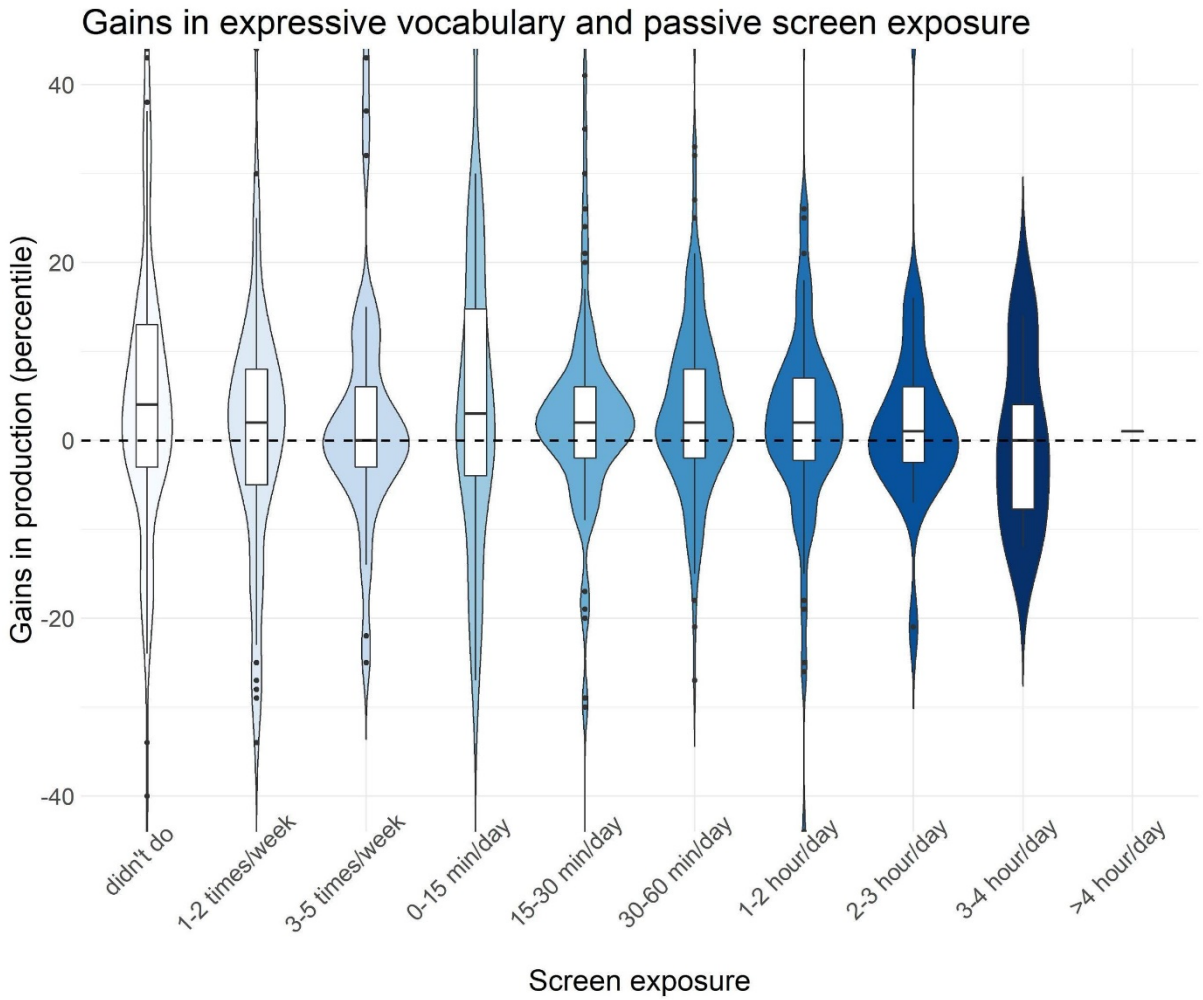


Figure 1. ***Violin plots of the gains in production (percentiles) for different amounts of reported child passive screen exposure. Boxplots display the first quartile and the third quartile, along with the median (the short horizontal bar). Gains of zero (dashed line) correspond to expected gains considering normative data.***

Maternal education, activities and gains in comprehension

A similar mixed-effect regression analysis was run on percentile gains in comprehension for children between 8- and 18-months of age (see Table 4) and on normalized raw gains in comprehension. Similar to the analyses on production, country and variation in maternal education by country were included as random factors.[6] Descriptive statistics for the activities and other variables used in the model can be found in

[6] Similar to the analyses of the production data, in order to address a potential issue of cryptic multiple testing raised by a reviewer, we performed, as recommended in Forstmeier & Schielzeth (2011), a full-null model comparison for both dependent variables (gains in percentiles and in normalized raw

the Supplementary Material 4.

Table 4. ***Fixed effects from the mixed-effect regression on the gains in comprehension. (left: percentiles with n = 330, right: raw scores with n = 585). p-values below .05 are marked in bold.***

	Gains in percentiles					Normalized gains in raw CDI scores				
	Est.	SE	t	df	p	Est.	SE	t	df	p
(Intercept)	**6.45**	**2.37**	**2.72**	**15.2**	**.02**	-3.65	7.12	-.51	42.9	.61
Maternal education	-.68	.89	-.76	9.0	.47	-.26	2.38	-.11	2.9	.92
Book reading	**1.48**	**.57**	**2.59**	**316.0**	**.01**	**3.55**	**1.06**	**3.35**	**544.2**	**.00**
Structured caregiver-child games	-.00	.45	-.00	312.9	1.00	1.17	.79	1.48	538.9	.14
Passive screen exposure	.03	.38	.07	268.7	.94	-.04	.78	-.05	538.8	.96
Outdoor activities	-.33	.31	-1.06	296.6	.29	-.38	.56	-.68	541.3	.50
Digital games	.45	.96	.46	311.7	.64	1.37	2.10	.65	526.8	.51
Free play w. child	.03	.42	.06	314.8	.95	-.78	.78	-1.01	534.5	.31
Singing	-.77	.47	-1.63	317.8	.10	-.44	.88	-.50	538.0	.62
Speaking	-.21	.36	.57	283.8	.57	-.34	.64	-.53	529.5	.59
Free play no adult	**-.80**	**.40**	**-2.01**	**311.4**	**.05**	-.67	.71	-.94	532.9	.35
Time gap	-.00	.05	-.07	102.8	.95	**.73**	**.09**	**7.89**	**149.7**	**.00**
Daycare before (yes)	-.93	1.71	-.54	313.3	.59	1.49	3.41	.44	482.7	.66
Gender (m)	-2.20	1.45	-1.51	311.4	.13	-4.95	2.69	-1.84	530.1	.07
Age (T1)	.02	.01	1.53	297.8	.13	-.01	.02	-.39	540.5	.70

Note. all numeric predictors were mean-centered in the analyses; p-values were calculated using Satterthwaite df.

CDI scores), where the full model contained all the factors included in the main model and the null model excluded the activities examined in the paper. The results of the full/null comparison revealed a significant difference between the two models in both gains in percentiles ($x^2 = 17.3$, $df = 9$, $p = .044$) and in normalized raw CDI scores ($x^2 = 19.8$, $df = 9$, $p = .019$), suggesting that the activities caregivers engaged their children with significantly improved the fit of the null model.

In both analyses, the time spent on shared book reading significantly correlated with gains in receptive vocabulary. As seen in Figure 2, children whose caregivers read 2-3 hours a day to them were reported to have the largest gains in receptive vocabulary size relative to the normative (age-matched) data. Yet, it is noteworthy that even participants with moderate exposure to books (more than 15 minutes per day) were reported to have gained more words than expected considering the (age-matched) norms.[7]

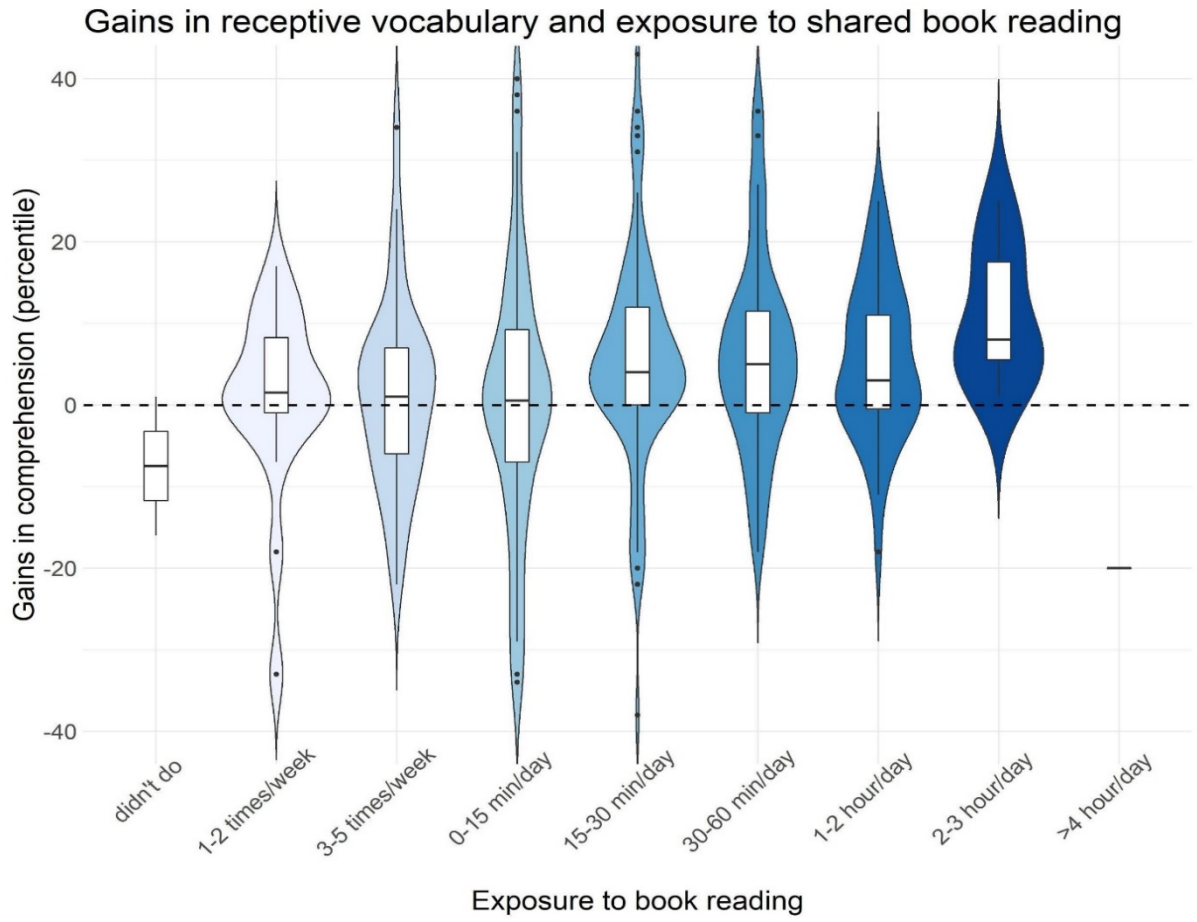


Figure 2. ***Gains in receptive vocabulary (in percentiles) for different amounts of reported shared book reading time. Gains of zero (dashed line) correspond to expected gains considering normative data.***

[7] As preregistered, similar to the analyses on production, we re-analyzed the data when >95% and <5% percentiles were excluded to check whether the model outcomes were impacted by these extreme values; the significant intercept and main positive effect of book reading remained significant (see details on OSF).

The quality of book reading, however, did not robustly correlate with gains in vocabularies, i.e., not across both measures of gains (see Analyses.Rmd on https://osf.io/ty9mn/ for the full analysis).

Similar to our analysis of production scores, analysis of caregiver reports suggested that young children (at the reference level of mean-centered age) gained more words in their receptive vocabularies during lockdown, i.e., daycare closure, when compared to the (age-matched) normative data (see Table 4 – the intercept is significantly above zero in the analysis on percentiles). A Wilcoxon signed-rank test with continuity correction found no evidence for a difference in children's reported vocabularies relative to normative data as children entered lockdown, at T1 (p = .9, Q1 = 23, median = 50, Q3 = 76), but found larger vocabularies relative to normative data at the end of lockdown (p = .01, Q1 = 29, median = 56, Q3 = 79). As indicated by the significant intercept, a one sample t-test in percentile gains between T1 and T2 revealed that, according to caregiver reports, children gained an average of 3.8 percentiles by T2 (95% CI [2.3, 5.2]; $t(317) = 5.0$, $p < .001$, $d = 0.28$)

A strong effect of time gap was also reported for the normalized gains in raw CDI scores, i.e., caregivers' vocabulary reports suggested that their children gained words throughout the lockdown. The additional effect of time spent playing without an adult in the percentile analysis did not replicate across analyses and will not be discussed further. There were no significant associations between children's gender or age and vocabulary development.

Maternal education and Vocabulary at T1

To estimate the extent to which maternal education was associated with expressive and receptive vocabulary at T1, in percentiles[8], we fitted two generalized linear mixed models with beta error structure and logit link function (McCullagh & Nelder, 1989; Bolker, 2008) using *glmmTMB* (Brooks, 2017). We fitted models with beta error structure due to issues with the homogeneity and normality of the residuals in the preregistered Gaussian model. The model revealed no effect of maternal education in either production ($\beta = 0.073$, $SE = 0.046$, $x^2 = 2.28$, $df = 1$, $p = .131$), or comprehension ($\beta = 0.041$, $SE = 0.058$, $x^2 = 0.501$, $df = 1$, $p = .479$, see Supplementary Material 5 for the full analysis). There were no significant associations between children's gender and receptive vocabulary at T1.

[8] Given that raw CDI sizes varied considerably across languages/tools (as number of items varied considerably across tools), correlated with age and we had wide variations in participants' ages across instruments, it was not possible to perform those analyses on raw CDI scores.

Discussion

Three findings stand out from the reported analyses. First, children who had less passive screen exposure during lockdown showed larger gains in their expressive, but not receptive, vocabulary size. Second, children whose caregivers read more to them during lockdown showed larger gains in their receptive, but not expressive, vocabulary size. Third, overall, based on caregivers' reports, children's receptive and expressive vocabularies showed larger increases during lockdown relative to their pre-lockdown, age-matched peers, i.e., using normative data collected pre-lockdown. We discuss these and other reported findings as well as provide potential explanations for these effects.

First, children who had more passive screen exposure during lockdown were reported to have lower gains in expressive vocabulary size (see Figure 1). Children who had no passive exposure to screens showed modest gains in expressive vocabulary relative to their pre-lockdown peers and smaller gains with increasing exposure to screens. There was no influence of passive screen exposure on children's receptive vocabulary across analyses. This differential association between screen exposure on receptive and expressive vocabulary size aligns with recent results in toddlers (Dynia et al., 2021). We suggest that the negative association between expressive vocabulary size and screen consumption may be explained by the fact that there is no requirement to respond to asynchronous digital content. This, in turn, may lead to longer stretches where children are not actively engaged in interacting with others, thereby providing them with little opportunity to expand their productive repertoire. In other words, digital media exposure may have an "opportunity cost" in that it takes time away from other interactions where children may have more opportunities to expand their expressive vocabulary. We did not collect information on the context of screen exposure, yet, recent research suggests that the context in which children are exposed to TV (e.g., during family meals, free day time, etc.) can have differential effects on language development (Martinot et al., 2021). A spin-off project on digital exposure provides more detail on digital practices in children and parents during the first covid lockdown (Bergmann et al., in press).

Second, we found that shared book reading explained more of the variance in gains in receptive vocabulary than any of the other examined activities (c.f. Montag et al., 2018). As shown in Figure 2, children whose caregivers did not engage in shared book reading at all were reported to have lower receptive vocabulary gains relative to pre-lockdown age-matched peers, whereas children whose caregivers engaged in more than 15-30 minutes of shared book reading per day were reported to have an increase in receptive vocabulary relative to pre-lockdown age-matched peers. There was no

similarly consistent association between shared book reading and children's expressive vocabulary size[9], nor between the quality of shared book reading and children's expressive or receptive vocabulary size. Our results highlight the association between book reading and some aspects of children's language development. Indeed, shared book reading includes more referential language than other routines and activities (Tamis-LeMonda et al., 2019); presents the child with higher frequencies of rare words than in everyday conversation (Montag et al., 2018) and allows children to explore words and worlds beyond the here and now.

It is noteworthy that reported receptive and expressive vocabulary growth during lockdown outpaced vocabulary growth in normative age-matched peers. There were no differences in the vocabulary increase between those infants who attended a daycare before the lockdown and who did not. While we did not predict such a lockdown boost, we suggest, post-hoc, alternative explanations for this finding. First, we may, perchance, be tapping into a demographic which differs from the sample used to calculate vocabulary norms. We suggest this to be unlikely given that we found no evidence that vocabulary sizes at T1 in our sample differed from normative data, nor did we find substantial differences in the distribution of maternal education in our sample and the one used to derive the vocabulary norms for the countries for which these data were available (see Supplementary Material 3). Second, many caregivers were working from home during lockdown and were with their child for longer stretches during the day relative to pre-lockdown. Thus, they had more opportunity to assess their child's development and might have been more aware of the words their child understood and produced, leading to more complete responding on the parent report forms we used and, hence, higher CDI scores. Third, social contact restrictions and closing of child-care facilities may have led to increased family and quality time between caregivers and children, providing them with more opportunities for activities that boost vocabulary knowledge, e.g., shared book reading. We are currently unable to disentangle the latter two interpretations of our findings and advocate caution in interpreting this lockdown boost in receptive vocabulary growth. Yet from a broader perspective these two interpretations need not be mutually exclusive: greater knowledge of children's vocabulary may allow caregivers to fine-tune the type and amount of input they provide to their child, in turn potentially leading to better outcomes (Fusaroli et al., 2019). Equally, children who showed greater improvements verbally may also have elicited particular interactions with their parents, e.g., increased amounts of time spent on shared book reading and less screen exposure. Other factors that might have modulated the role of activities are the household structure, the presence (and, if so, the number) of siblings, which is examined in a separate spin-off project, and the circumstances of data collection. Given that the data were collected during the first COVID-19 lockdown, it is possible that parents' engagement

[9] The relationship between book reading and gains in expressive vocabulary was only revealed for the normalized gains in vocabulary.

in the study was affected by the ongoing pandemic and differed from the non-COVID-19 times, when parents have other demands on their time and attention and feel less stressed. Recent studies reported that the pandemic affected mothers in particular (Langin, 2021), as mothers spent more time to take care about the child and the household than fathers, and mothers' experience of pandemic (not measured in the current study) might have influenced their behavior and responsiveness (Evans et al., 2021).

Importantly, children entered the lockdown with a range of vocabulary sizes and had been exposed to learning environments differing in quality prior to daycare closure. The associations between shared book reading, screen time and receptive and expressive vocabulary development, respectively, reported above are considerable, as they capture associations between momentary modulations in the child's learning environment (over an average of just 41 days) and vocabulary development. This is especially so, given recent findings suggesting that parental input shapes children's language skills even after controlling for potential genetic confounds (Coffey et al., 2021). Other activities (outdoor activities, caregiver-child interactions/games), that did not predict gains in receptive and expressive vocabulary size, contributed to other aspects of the child's development, such as the child's well-being during the lockdown (currently being investigated in a separate spin-off project). In contrast to book reading and screen exposure – the two activities that have been systematically analyzed in recent child development research - there are no standardized questionnaires that cover the wide spectrum of languages used in the present study, to examine, retrospectively, child-parent engagement across the wider set of activities used in the current study, e.g., singing, outdoor activities. Therefore, the lack of a significant effect of other activities on vocabulary gains might be attributed, to the lack of salience of other activities to parents, to unknown psychometric properties of reports associated with some activities (e.g., most infants did not use digital games in our study), or to limited reliability when parents are asked to recall past activities (Nivison et al., 2021). However, the analysis, over the same cohort, of the impact of activities on a child's well-being - the focus of a separate contribution (see https://osf.io/ns6gh/?view_only=bee2c0f1686542e9b006ea04e36f0c88)- suggests that parental reports can be used across a range of activities, and that varying activities might have differential effects on child's language development and well-being.

Contrary to our hypothesis, maternal education did not correlate with receptive or expressive vocabulary growth during lockdown or vocabulary size at the onset of lockdown. Note that the absence of an effect of maternal education on gains in receptive or expressive vocabulary size should be taken with caution, as there were relatively few participants with the maternal education lower than a Bachelor degree, which was level 4 on a scale from 1 to 6 in our study (14% of the comprehension data and 10% of the production data) and few participants with the high-school education level only, which was level 2 on our scale (5% of the comprehension data and 3% of the production data). Although the proportion of mothers with low education level in the

current sample was comparable to that reported in the normative data for some of the countries in wordbank.stanford.edu (see Supplementary Material 3), research on a sample with a more homogeneous distribution of maternal education is required to further address this question. Therefore, the extent to which these findings generalize to families from lower socioeconomic backgrounds (as indexed by lower education level in the current study) and less industrialized countries, who were hit hardest by the pandemic, remains uncertain. Nonetheless, the absence of the effect of maternal education is consistent with the modest effects of maternal education on vocabulary reported in data from Wordbank (excluding the USA; Frank et al., 2021) particularly in children under 24 months, especially since a large percentage of the current sample involved children below this age (68%). However, maternal education did correlate positively with time spent on shared book reading, and negatively with time the child spent with digital media. Thus, while there were differences in the activities that caregivers with differing levels of educational attainment engaged in with their child (Entwisle et al., 2001; Pace et al., 2017), our results suggest that the activities that caregivers engaged in with their children, rather than caregivers' educational attainment, correlated with children's receptive and expressive vocabulary development during lockdown. The conjunction of these results highlights some of the pathways through which maternal education (as a proxy for SES) may explain variability in vocabulary development in other studies (Fernald et al., 2013; Pace et al., 2017, Rowe, 2018).

Conclusion

This large-scale multinational study (1742 participants, 13 countries) offers a unique window into associations between features of the home environment and children's longitudinal receptive and expressive vocabulary development. Taken together, the results suggest, that in our sample, caregiver education, children's age or sex were not associated with children's receptive and expressive vocabulary development as much as some of the activities that caregivers reported undertaking with their children.

In particular, the frequency and duration of shared book reading and screen exposure were related to respective receptive and expressive vocabulary gains in lockdown – children whose caregivers read more to them and who had less passive screen exposure showed larger receptive and expressive vocabulary gains, respectively, – and that children's reported receptive and expressive vocabulary development was boosted compared to pre-pandemic CDI norms.

References

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking Quality and Quantity of Parental Linguistic Input to Child Language Skills: A Meta-Analysis. *Child Development*, 92(1). https://doi.org/10.1111/cdev.13508.
Baayen, R.H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics. Cambridge: Cambridge University Press.

Barr, D.J., Levy, R., Scheepers, C. & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014. http://dx.doi.org/10.18637/jss.v067.i01

Bergmann, C., Dimitrova, N., Alaslani, K., Almohammadi, A., Alroqi, H., Aussems, S., Barokova, M., Davies, C., Gonzalez-Gomez, N., Gibson, S. P., Havron, N., Horowitz-Kraus, T., Kanero, J., Kartushina, N., Keller, C., Mayor, J., Mundry, R., Shinskey, J. L., & Mani, N. (in press). Young children's screen time during the first COVID-19 lockdown in 12 countries. *Scientific reports* https://doi.org/10.31219/osf.io/p5gm4

Bolker, B. M. (2008). *Ecological models and data in R*. Princeton University Press. ISBN: 9780691125220

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, 9(2), 378-400. https://doi.org/10.32614/RJ-2017-066

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, 110(28), 11278-11283. https://doi.org/10.1073/pnas.1309518110

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development*, 91(5), 1819-1835. https://doi.org/10.1111/cdev.13349

Coffey, J. R., Shafto, C. L., Geren, J. C., & Snedeker, J. (2021). The effects of maternal input on language in the absence of genetic confounds: Vocabulary development in internationally adopted children. *Child Development*, 93(1).

https://doi.org/10.1111/cdev.13688

Davis-Kean, P. E., Tighe, L. A., & Waters, N. E. (2021). The Role of Parent Educational Attainment in Parenting and Children's Development. *Current Directions in Psychological Science,* 30(2), 186–192. https://doi.org/10.1177/0963721421993116

deMayo, B., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C. F., Frank, M. C., & Marchman, V. A. (2021). Web-CDI: A system for online administration of the MacArthurBates Communicative Development Inventories. *Language Development Research,* 1(1), p 55-98. https://doi.org/10.34758/kr8e-w591

Dynia, J. M., Dore, R. A., Bates, R. A., & Justice, L. M. (2021). Media exposure and language for toddlers from low-income homes. *Infant Behavior and Development,* 63, 101542. https://doi.org/10.1016/j.infbeh.2021.101542

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2001). Keep the faucet flowing summer learning and home environment. *American Educator,* 25(3), 10-15. ISSN: ISSN-0148-432X.

Evans, D. K., Jakiela, P., & Knauer, H. A. (2021). The impact of early childhood interventions on mothers. *Science,* 372(6544), 794-796. https://doi.org/10.1126/science.abg0132

Fenson, L. (2007). *MacArthur-Bates communicative development inventories.* Baltimore, MD: Paul H. Brookes Publishing Company. ISBN 13: 978-1557668882

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science,* 16(2), 234-248. https://doi.org/10.1111/desc.12019

Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word learning: A meta-analysis. *Developmental Psychology,* 54(7), 1334. https://psycnet.apa.org/doi/10.1037/dev0000512

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology,* 65(1), 47–55. https://doi.org/10.1007/s00265-010-1038-5

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language,* 44(3), 677. https://doi.org/10.1017/S0305000916000209

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press. https://doi.org/10.1162/opmi_a_00026

Fusaroli, R., Weed, E., Fein, D., & Naigles, L. (2019). Hearing me hearing you: Reciprocal effects between child and parent language in autism and typical development. *Cognition*, 183, 1-18. https://doi.org/10.1016/j.cognition.2018.10.022

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., & Singer, D. G. (2009). *A mandate for playful learning in preschool: Presenting the evidence*. Oxford University Press, USA. Print ISBN-13: 9780195382716

Han, J., & Neuharth-Pritchett, S. (2015). Meaning-related and print-related interactions between preschoolers and parents during shared book reading and their associations with emergent literacy skills. *Journal of Research in Childhood Education*, 29(4), 528-550. https://doi.org/10.1080/02568543.2015.1073819

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing. ISBN: 978-1-55766-197-5

Langin, K. (2021). Pandemic hit academic mothers hard, data show. *Science*, 371(6530), p 660. https://doi.org/10.1126/science.371.6530.660

Long, J. A. (2020). Package 'jtools'.

Malin, J. L., Cabrera, N. J., & Rowe, M. L. (2014). Low-income minority mothers' and fathers' reading and children's interest: Longitudinal contributions to children's receptive vocabulary skills. *Early Childhood Research Quarterly*, 29(4), 425-432. https://doi.org/10.1016/j.ecresq.2014.04.010

Mayor, J., & Mani, N. (2019). A short version of the MacArthur–Bates Communicative Development Inventories with high validity. *Behavior Research Methods*, 51(5), 2248-2255. https://doi.org/10.3758/s13428-018-1146-0

Martinot, P., Bernard, J. Y., Peyre, H., De Agostini, M., Forhan, A., Charles, M.-A., Plancoulaine, S., & Heude, B. (2021). Exposure to screens and children's language development in the EDEN mother–child cohort. *Scientific Reports*, 11(1), 11863. https://doi.org/10.1038/s41598-021-90867-3

McCullagh, P., & Nelder, J. A. (1989). Monographs on statistics and applied probability. *Generalized linear models*, 37. ISBN-13: 978-0412317606

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, 26(9), 1489-1496. https://doi.org/10.1177%2F0956797615594361

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive Science*, 42, 375-412. https://doi.org/10.1111/cogs.12592

Nivison, M. D., Vandell, D. L., Booth-LaForce, C., & Roisman, G. I. (2021). Convergent and Discriminant Validity of Retrospective Assessments of the Quality of Childhood Parenting: Prospective Evidence From Infancy to Age 26 Years. Psychological Science, 32(5), 721-734.

Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying pathways between socioeconomic status and language development. *Annual Review of Linguistics*, 3, 285-308. https://doi.org/10.1146/annurev-linguistics-011516-034226

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, 76(4), 763-782. https://doi.org/10.1111/1467-8624.00498-i1

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Rowe, M. L. (2018). Understanding socioeconomic differences in parents' speech to children. *Child Development Perspectives*, 12(2), 122-127. https://doi.org/10.1111/cdep.12271

Schielzeth, H. & Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavoral Ecology*, 20, 416-420.

Shahaeian, A., Wang, C., Tucker-Drob, E., Geiger, V., Bus, A. G., & Harrison, L. J. (2018). Early shared reading, socioeconomic status, and children's cognitive and school competencies: Six years of longitudinal evidence. *Scientific Studies of Reading*, 22(6), 485-502. https://doi.org/10.1080/10888438.2018.1482901

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2019). Routine language: Speech directed to infants during home activities. *Child Developmen*t, 90(6), 2135-2152 https://doi.org/10.1111/cdev.13089

UNESCO (2020). COVID-19 educational disruption and response. UNESCO.
van den Heuvel, M., Ma, J., Borkhoff, C. M., Koroshegyi, C., Dai, D. W., Parkin, P.

C., … & Birken, C. S. (2019). Mobile media device use is associated with expressive language delay in 18-month-old children. *Journal of Developmental and Behavioral Pediatrics*, 40(2), 99. https://doi.org/10.1097/DBP.0000000000000630

van den Heuvel, M., Ma, J., Borkhoff, C. M., Koroshegyi, C., Dai, D. W. H., Parkin, P. C., Maguire, J. L., & Birken, C. S. (2019). Mobile Media Device Use is Associated with Expressive Language Delay in 18-Month-Old Children. *Journal of Developmental and Behavioral Pediatrics*, 40(2), 99–104. https://doi.org/10.1097/DBP.0000000000000630

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly*, 37, 39-57. https://doi.org/10.1016/j.ecresq.2016.04.003

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24(11), 2143-2152. https://doi.org/10.1177/0956797613488145

Williams, K. E., Barrett, M. S., Welch, G. F., Abad, V., & Broughton, M. (2015). Associations between early shared music activities in the home and later child outcomes: Findings from the Longitudinal Study of Australian Children. *Early Childhood Research Quarterly*, 31, 113-124. https://doi.org/10.1016/j.ecresq.2015.01.004

Zimmerman, F. J., Christakis, D. A., & Meltzoff, A. N. (2007). Associations between media viewing and language development in children under age 2 years. *The Journal of Pediatrics*, 151(4), 364-368. https://doi.org/10.1016/j.jpeds.2007.04.071

Data, code and materials availability statement

Data, code and materials that support the findings of this study are openly available on the OSF repository at https://osf.io/ty9mn/

Ethics statement

The research project was approved by the Norwegian Center for Research Data REF536895 and by the ethics committee of the Department of Psychology at the University of Oslo. Collaborating labs obtained ethical approval from their institutions.

Authorship and Contribution Statement

NK and JM conceptualized and designed the study. NK and JM preregistered the original analyses for the Norwegian data. NK, JM, and NM created material for the study. NK, NM, AEA, KA, NA, AA, HA, LA, EA, SA, MB, MDB, CB, CC, SC, ADC, ND,

AD, RF, CF, AF, MG, HGS, SG, KG, NGG, EH, EEH, NH, JH, CH, THK, MK, JK, CK, GK, CL, RL, MŁ, KM, KN, NO, LP, CR, DSO, JS, AV, KV, MZ and JM contributed to data collection. NK, JM, CL, CB, SA, MB, GK, JK, NA, NH, NM and MK preregistered planned analyses for the full sample. NK, LM and JM processed the data, and NK led the data analysis with JM. NK, NM and JM interpreted the data with input from SA, CB, NH, JK and MF. NK, NM and JM wrote the manuscript. NK, NM, JM, CL, CB, SA, MB, ND, THK, JK, MB, NA, RL, HA, AA, KA, NGG, SG, LP, SC, MK, CR, MF, EH reviewed and revised the manuscript. All authors proof-read and approved the final version of the manuscript for submission.

Acknowledgments

We thank Roger Mundry for his analysis of the role of SES on vocabularies at T1. We are grateful to families who took part in the study in these challenging times. We would like to thank reviewers for their insightful comments and suggestions. Funding for Polish sample: grant from National Science Center NCN (Poland), no 2018/31/B/HS6/03916. NK was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme [project number 223265].

Appendices

Supplementary Material 1

Table 1. ***Sample sizes for participating labs, for production data.***

Labid	Language	Country	T1 sample	T2 sample	Final T1-T2 sample	Average T1-T2 gap (days)
kau-cll	Arabic	Saudi Arabia	336	171	90	73
nbu	Bulgarian	Bulgaria	69	18	0	
brc-nijmegen	Dutch	The Nether-lands	26	25	20	39
brookes	English	UK	565	341	292	74
cogdevlabbyu	English	USA	93	89	39	23
clcu	English	UK	123	56	40	35
dsc	English	USA	32	14	5	86
ilpll	English	USA	263	115	49	73
ldl	English	Canada	63	29	17	33
Louisville	English	USA	62	nc	na	
owll	English	USA	10	nc	na	
rhul_baby_lab	English	UK	55	34	25	25
unlv	English	USA	56	27	0	
paris_team	French	France	654	535	466	28
goe	German	Germany	84	69	37	63
HaifaUniv	Hebrew	Israel	343	103	61	26
technion_il	Hebrew	Israel	335	164	111	37
babyling	Norwegian	Norway	786	182	173	20
multilada	Polish	Poland	670	246	223	27
hetsl	French	Switzerland	nc	400	ca	
msu	Russian	Russia	255	24	17	41
bcbl	Spanish	Spain	157	131	37	65
mltlab	Turkish	Turkey	57	57	40	31
		Total	5094	2830	1742	41

nc - data not collected
na - does not apply, giving that data in one sample was missing
ca - due to the lack of child's exact age, Swiss final data (n = 290) was used in the analyses of the relationship between maternal education and activities (cf project's OSF)
Note. Final T1-T2 sample contains data points that have passed the inclusion criteria after the merge of the matching T1 and T2 questionnaires.

Supplementary Material 2

Figure 1. ***Non-normalized (top) and normalized (bottom) gains in comprehension vocabulary as a function of the adjusted CDI score at T1 for the CDI tools Words and Gestures (wg) and Words and Sentences (ws). See Analyses_2.Rmd code on https://osf.io/ty9mn/ for data on production.***

Supplementary Material 3

We report below (Table 2) the fraction of participants having completed primary (level 1) and "some secondary" (level 2) education in the current sample, as well as in the sample used to derive vocabulary norms (from WordBank). We restricted our comparison to the handful of instruments for which we have maternal education information in both samples, as well as having commensurate measures of maternal education.

Table 2. Percentage of participants in the first two levels of maternal education scale (primary, and some secondary), for the norming sample (WordBank) and our sample. Differences in the maternal education between the Wordbank sample and the German and Spanish samples in the current study are likely attributed to smaller sample sizes in these two countries in our study.

Instrument (CDI)	Percentage of participants on WordBank	Percentage of participants in our sample (and sample size)
American English CDI	5.3%	5.5% (110)
Norwegian CDI	5.0%	4.0% (173)
French CDI	0%	1.1% (466)
German CDI	37.1%	13.5% (37)
Spanish CDI	5.4%	8.1% (37)

Supplementary Material 4

Table 3. ***Descriptive statistics of the variables used in the analyses of the production data (in percentiles).***

	Maternal education	Age at T1 (in days)	Book reading	Structured caregiver-child	Passive screen exposure	Outdoor activities	Digital games	Free play w. child	Singing	Speaking	Free play no adult	Time gap
Mean	4.5	588.9	4.1	2.5	3.3	4.4	0.5	5.8	3.7	5.9	5.2	35.8
SD	0.9	195.8	1.6	1.9	2.4	2.7	1.3	1.9	1.7	2.1	1.9	21.5
Min	1	245	0	0	0	0	0	0	0	0	0	4
Max	6	1075	9	8	9	9	7	9	9	9	9	111

Table 4. ***Descriptive statistics of the variables used in the analyses of the comprehension data (in percentiles).***

	Maternal education	Age at T1 (in days)	Book reading	Structured caregiver-child	Passive screen exposure	Outdoor activities	Digital games	Free play w. child	Singing	Speaking	Free play no adult	Time gap
Mean	4.4	418.6	3.9	2.0	2.4	3.8	0.2	5.9	3.9	6.0	4.9	40.3
SD	0.9	73.9	1.5	1.9	2.3	2.6	0.8	1.8	1.7	2.3	1.9	22.9
Min	1	245	0	0	0	0	0	0	0	0	0	5
Max	6	566	9	7	8	9	5	9	9	9	9	111

Supplementary Material 5

Impact of SES on Vocabulary at T1

To estimate the extent to which language comprehension and production depended on maternal education we fitted two Generalized Linear Mixed Models (GLMM; Baayen 2008) with beta error structure and logit link function (McCullagh & Nelder 1989; Bolker 2008). We used a beta rather than a Gaussian error function since the residuals of the Gaussian model were neither normally distributed nor homogeneous. Both models were identical in their fixed and random effects: As fixed effects, we included maternal education while controlling for sex, i.e., two fixed factors. We included random intercepts of country and random slopes of both predictors within country (cf Schielzeth & Forstmeier 2009; Barr et al. 2013). We excluded parameters for the correlations among the random intercept and slopes due to model convergence issues.

Maternal education was z-transformed (M=0, SD=1) to ease model convergence and the random effect of sex was manually dummy coded and centered. We fitted the model in R (version 4.0.3; R Core Team 2020) using the function glmmTMB (version 1.0.2.1; Brooks 2017). We determined the significance of individual fixed effects by comparing the respective full model with reduced models lacking them one at a time, utilizing likelihood ratio tests (Dobson 2002). We determined confidence intervals of model estimates by means of a parametric bootstrap (function simulate of the package glmmTMB) and estimated model stability by dropping countries one at a time and comparing estimates of models fitted to the respective subsets of the data to those obtained for the full data set. This revealed both models to be of moderate to good stability (see results). Neither of the two models was overdispersed (dispersion parameters; comprehension model: 1.00; production model: 1.048). The samples analysed for the two models comprised a total of 352 children from eight countries (comprehension model) and a total of 729 children from nine countries (production model).

As can be seen in Tables 1 and 2, neither maternal education nor sex were significant in either of the two models. However, sex was only marginally non-significant in the production model and all model estimates had the hypothesized sign (Tables 1 and 2).

Table 5. ***Results of the comprehension model (estimates together with standard errors, confidence limits, significance tests, and range of estimates (min, max) when dropping countries one at a time).***

term	Estimate	SE	lower Cl	upper Cl	x^2	df	P	min	max
Intercept	0.092	0.164	-0.227	0.429			[1]	-0.096	0.185
mat. educ[2]	0.041	0.058	-0.075	0.162	0.501	1	0.479	0.031	0.083
sex[3]	-0.146	0.151	-0.454	0.136	0.840	1	0.360	-0.254	-0.040

[1] not indicated because of having a very limited interpretation
[2] z-transformed to a mean of zero and a standard deviation (sd) of one; mean and sd of the original variable were 4.375 and 0.944, respectively
[3] dummy coded with female being the reference category

Table 6. ***Results of the production model (estimates together with standard errors, confidence limits, significance tests, and range of estimates (min, max) when dropping countries one at a time).***

term	Estimate	SE	lower Cl	upper Cl	x^2	df	P	min	max
Intercept	0.053	0.115	-0.169	0.296			-1.000	-0.018	0.121
mat. educ[2]	0.073	0.046	-0.017	0.162	2.285	1	0.131	0.040	0.105
gender[3]	-0.268	0.130	-0.520	-0.013	3.061	1	0.080	-0.349	-0.178

[1] not indicated because of having a very limited interpretation
[2] z-transformed to a mean of zero and a standard deviation (sd) of one; mean and sd of the original variable were 4.505 and 0.882, respectively
[3] dummy coded with female being the reference category

License

It's your turn: The dynamics of conversational turn-taking in father-child and mother-child interaction

Linda Kelly
Elizabeth Nixon
Jean Quigley
School of Psychology, Trinity College Dublin, Ireland

Abstract: The aim of this study was to elucidate the interactive and temporal features of conversational turn-taking during father-child and mother-child play and investigate associations with children's cognitive and language abilities. Eighty typically developing two-year-olds (M = 24.06 months, SD = 1.39) and their biological mothers and fathers took part in the current study which consisted of a single visit to an Infant and Child Lab. Parent-child conversational turn-taking was measured from dyadic structured play interactions (160 dyads in total), as well as parents' verbal turn-taking behaviours including length of turn, questions, and contingent responsiveness. Child language and cognitive skills were directly assessed using standardised measures. Results indicated that there was greater balance in conversational turn-taking during father-child play. However, mothers were more responsive to their child's vocalisations during interaction. Mothers' and fathers' use of questions effectively scaffolded children's participation in conversation. Finally, controlling for mother-child conversational turn-taking, father-child conversational turn-taking did not account for any unique variance in child cognitive skills. Regression analyses failed to demonstrate associations between parent-child conversational turn-taking and child language skills. These findings present new insights into the dynamics of mother-child and father-child conversational turn-taking during play as well as the nature of the contribution of father-child linguistic exchanges to child development.

Keywords: fathers; child-directed speech; conversational turn-taking; language development; cognitive development.

Corresponding author(s): Linda Kelly, School of Psychology, Trinity College Dublin, College Green, Dublin 2, Ireland, D02PN40. Email: kellyl11@tcd.ie.

ORCID ID(s): Linda Kelly https://orcid.org/0000-0002-7687-9248; Elizabeth Nixon https://orcid.org/0000-0001-8746-4390; Jean Quigley https://orcid.org/0000-0003-0469-5199

Citation: Kelly, L., Nixon, E., & Quigley, J. (2022). It's your turn: The dynamics of conversational turn-taking in father-child and mother-child interaction. *Language Development Research, 2*(1), 37—68. https://doi.org/10.34842/840g-2297

Introduction

Socio-cultural and social-interactionist theories of development emphasise how variation in the quality of social-communicative interactions between parents and their children contribute meaningfully to child development (Bruner, 1981; Snow, 1977; Vygotsky, 1978). Child-directed speech (CDS) is an important communicative tool which parents use, seemingly intuitively, that serves a variety of important functions for the developing child (Fernald, 1989; Rowe, 2012; Werker & McLeod, 1989). Traditionally, research on children's early linguistic environment has focussed on the mother-child dyad, but we know that fathers contribute in important and unique ways to child development (Cabrera et al., 2014; Lamb & Lewis, 2010). Including both mothers and fathers in research is valuable in providing a closer approximation of the ecology of the developing child and the range of factors which shape their development.

Apart from the lexical and syntactic features of CDS, studies have demonstrated the importance of pragmatic dimensions of parental input during toddlerhood (Rowe & Snow, 2020). According to Bruner (1983), children's development relies on more than exposure to language input, and it is important to emphasise the interactive component of parent-child communication. In particular, recent literature has turned its focus to the importance of conversational turn-taking in parent-child interaction for child development (e.g., Donnelly & Kidd, 2021; Gilkerson et al., 2017; Gómez & Strasser, 2021; Romeo et al., 2018). However, little research thus far has specifically examined conversational turn-taking in father-child interaction. The aim of the current study was therefore to examine conversational exchanges in mother-child and father-child interaction. Furthermore, in order to better understand the dynamics of these communicative exchanges, the present study sought to decompose the construct of conversational turn-taking and examine how mother-child and father-child interactive verbal behaviours support young children's engagement in back-and-forth exchanges. Lastly, this study investigated concurrent associations between father-child conversational turn-taking and children's language and cognitive abilities.

Conversational Turn-Taking and Child Development

Newborn infants show an early propensity for social interaction and the behaviours of both infants and their parents are intent on promoting and maintaining proximity with one another (Bowlby, 1969). Before they learn to speak, infants engage in episodes of joint attention with their parents and communicate using behaviours such as vocalisations and facial expressions. These behaviours are highly contingent upon and synchronised with those of their parents (Trevarthen & Aitken, 2001). Bateson (1979) termed these pre-linguistic interactions between infant and parents "proto-conversations" and described these exchanges as the early precursors of conversation and turn-taking. The "conversational duet" in which parent and child are jointly engaged in interaction is also considered an important foundation for child language

and socio-cognitive development (Bruner, 1983; Hirsh-Pasek et al., 2015). This aligns with transactional models which frame the development of the individual as arising from dynamic bidirectional interactions between the child and their environment (Sameroff, 2009).

The literature proposes several pathways by which parent-child conversational turn-taking may support children's development. Back-and-forth verbal exchanges between parents and children may help caregivers gauge the developmental capacities of their child and pitch the complexity of their language input within the bounds of the child's zone of proximal development, maximising their learning potential (Vygotsky, 1978). Greater conversational turn-taking may be indicative of high levels of caregivers' responsiveness, which may explain how greater involvement in conversation drives child language learning (Zimmerman et al., 2009). Involvement in conversation also provides children the opportunity to practice their emerging language and cognitive skills and may support deeper engagement by the child with the linguistic structure of speech input (Romeo et al., 2018). Beyond exposure to language input, studying children's involvement in conversation provides an insight into the child's active role in their own development.

Research to date has demonstrated that during early childhood, conversational turn-taking in parent-child interaction may be a stronger predictor of child language and brain development than quantity of parental speech input (e.g., Gilkerson et al., 2017; Romeo et al., 2018; Zimmerman et al., 2009). Romeo and colleagues (2018) demonstrated that, controlling for quantity of speech input, conversational turn-taking between children aged 4–6 years and their parents was associated with children's brain activity and their verbal abilities. Longitudinal research has also demonstrated that controlling for quantity of input, more conversational turn-taking between parents and preschool aged children was associated with greater language abilities 18 months later (Zimmerman et al., 2009). In another longitudinal study, Gilkerson and colleagues (2017) examined conversational turn-taking between children aged 2–48 months of age and their caregivers at monthly intervals and observed associations with child language ability.

Gilkerson and colleagues (2018) also demonstrated that early conversational turn-taking predicted child IQ and verbal abilities 10 years later. The authors observed that conversational turn-taking between caregivers and their children which took place during the window of 18–24 months of age was particularly important for later child outcomes. Recently, Donnelly and Kidd (2021) demonstrated bidirectional associations between adult-child conversational turn-taking and children's vocabulary development between 9–24 months of age. Children become more proficient turn-takers as their language skills advanced, and at the same time conversation with caregivers emerged as an important context for children's language development (Donnelly & Kidd, 2021). Overall, the findings of these studies emphasise the importance of

studying the interactive components of children's early communicative environments.

Previous studies of conversational turn-taking are however subject to several limitations. Research to date has relied on data produced by The Language Environment Analysis (LENA) system, a widely used tool for measuring day-long recordings. Recent studies evaluating LENA suggest that, compared to human coders, this system may miss more instances of speech and is less effective in tagging speakers correctly (Cristia et al., 2020). A longitudinal study which compared LENA's adult-child conversational turn count to manually coded turn counts at five time points between 6–24 months of age also demonstrated that LENA overestimated turn counts across all age groups (Ferjan Ramírez et al., 2021). In addition, it can be unclear when using this tool as to whether the speech in the child's environment was directed towards the child or was merely overheard (Zimmerman et al., 2009).

Furthermore, LENA relies on counts of conversational turns in the child's interactive environment. This approach, however, fails to account for the distribution of conversational load across the interaction. A conversational turn begins when one interlocutor starts speaking and ends when the next speaker commences. One conversational turn can therefore consist of several utterances. Comparing both parent's and child's mean length of turn provides insight into how interlocutors share the burden of conversation within turn-taking episodes. Greater balance in turn-taking occurs when parent and child take turns of similar length and no one interlocutor is dominating the conversation. Equilibrium in turn-taking suggests that both interlocutors are actively verbally participating in conversation across the interaction and may be more effective in capturing children's engagement in conversation compared with conversational turn counts. Conversational balance is calculated by computing the ratio of each interlocutor's mean length of turn within a conversation (see Lloyd et al., 2001; McDonnell et al., 2003; Vaughan et al., 2015 for examples of other studies using this approach).

Examination of conversational balance provides insight into children's involvement in conversation but reveals little information with regards to the qualitative content of the conversations between parent and child and the turn-taking behaviours exhibited by parents which support children's participation in language interactions. If conversational turn-taking is an important aspect of the early interactive environment, as emerging research suggests, it is of interest to understand more clearly the dynamics of conversational turn-taking and the mechanisms through which it may support child language and cognitive development.

Finally, a key limitation of previous research is the lack of focus on father-child conversational turn-taking. Early father-child language exchanges have important implications for children's language and cognitive development (Rowe et al., 2017; Schwab,

et al., 2018), often beyond the influence of maternal CDS (Baker & Vernon-Feagans, 2015; Conica et al., 2020; Malin et al., 2014; Pancsofar & Vernon-Feagans, 2006; Pancsofar & Vernon-Feagans, 2010). Studies comparing mothers' and fathers' CDS during toddlerhood have, however, primarily focussed on the lexical and syntactic features of speech rather than the interactive elements of parent-child communication. This study therefore sought to profile both mother-child and father-child conversational turn-taking during play and examine how parents' interactive verbal behaviours support children's verbal participation in conversation as well as their language and cognitive development.

Dynamics of Turn-Taking during Parent-Child Conversation

Certain features of parents' speech and communication may serve to scaffold children's participation in conversation. The present study was concerned with elucidating whether certain interactive verbal behaviours produced by mothers and fathers were associated with greater balance in turn-taking in parent-child conversation. The units of turn-taking explored in the current study included parents' length of turn, questions posed by mothers and fathers, and parental contingent responsiveness.

Length of Turn

The first interactive verbal behaviour examined by the present study was parents' length of turn. As previously mentioned, one conversational turn can comprise multiple utterances. Longer turns may indicate that one interlocutor is dominating the language interaction. Parents who take longer turns may be providing fewer opportunities for their child to participate in conversation. Previous research has demonstrated that when parents decreased the length of turns they took, children's verbal participation in conversation increased (Brassart & Schelstraete, 2015; Girolametto, 1988). The literature suggests the CDS that mothers and fathers produce during interaction with their toddlers is comparable (Pancsofar & Vernon-Feagans, 2006; Rowe et al., 2004), therefore it was hypothesised that no significant differences between mothers' and fathers' length of turn would be observed. If greater conversational turn-taking is associated with better child language and cognitive scores, it was expected that parents' length of turn would be inversely related to child developmental abilities.

Parental Contingent Responsiveness

Another important aspect of back-and-forth exchanges is responsiveness. As young children develop greater competency as communicators, parents hold much of the responsibility for coordinating smooth verbal exchanges, and this is facilitated by responding contingently to the child's vocalisations (Rutter & Durkin, 1987). Conversational turn-taking may therefore be enhanced by sensitive and contingent responding to the child (Brassart & Schelstraete, 2015). Well-timed responses are typically

considered to occur within 2–5 seconds of a child's utterance (McGillion et al., 2013). Semantically contingent responding is also a prerequisite of successful verbal interaction (Bornstein et al., 2015) whereas parental utterances which fail to follow the child's focus of attention may be less useful in supporting children's engagement in conversation (Brassart & Schelstraete, 2015).

Research with mothers has consistently shown that responses which are well-timed and semantically related to the child's present focus of attention facilitate child language and cognitive development (Bornstein et al., 1999; Landry et al., 2000; Masur et al., 2005; Tamis- LeMonda et al., 2001; Tamis-LeMonda et al., 2014). Parental responsiveness in early infancy may serve to convey the role of language as a social-communicative device (Tamis-LeMonda et al., 2001). It may also help children to match labels to objects in the environment thereby supporting vocabulary development (Tamis-LeMonda & Bornstein, 2002). Furthermore, responsive caregiving may contribute to the child's emerging sense of their own impact on the world around them (Bornstein et al., 2015), perhaps furnishing them with an awareness of their own behaviour and capacity for regulation (Kopp, 1982). Compared to mothers, much less is known about fathers' responsiveness during parent-child interaction although research suggest that fathers' sensitivity to their children's cues is important for cognitive and language development (Tamis-LeMonda et al., 2004).

Questions

Another turn-taking behaviour studied in the literature is questions produced by mothers and fathers during interaction with their child. Locke (1996) suggested that while turn-taking with younger children is primarily supported by parents' contingent responsiveness, by age 24 months caregivers place more responsibility upon children to participate in conversation by asking questions. Previous studies suggest that fathers produce more conversation-eliciting speech such as *wh*-questions during interaction with their young children compared to mothers (Malin et al., 2014; Rowe et al., 2004) although others (e.g., Pancsofar & Vernon-Feagans, 2006) observed no difference. Conversation-eliciting speech is hypothesized to be a challenging feature of the child's communicative environment and has previously been demonstrated to support child verbal reasoning (Rowe et al., 2017) and language development (Leech et al., 2013). *Wh*-questions may require complex responses compared to yes/no questions and may therefore support children's development of language and reasoning skills (Rowe et al., 2017). It was also expected that a higher proportion of CDS in the form of questions posed by parents would encourage greater verbal participation of the child during interaction.

The Current Study

Research focussing solely on the role of mothers overlooks the rich ecology of the

developing child. This study sought to more comprehensively characterise the child's early linguistic environment by examining conversational turn-taking in father-child and mother-child interaction. The first aim of the current study was to present a profile of parents' interactive verbal behaviours produced during parent-child interaction and compare these between mothers and fathers. Given the absence of previous research comparing mother-child and father-child conversational turn-taking, no specific hypothesis was made in this regard. In relation to parents' interactive verbal behaviours, and in light of previous research, it may be expected that fathers would produce more *wh*-questions compared to mothers. On the other hand, previous research suggests that mothers may display more contingent responsiveness in interaction compared to fathers (e.g., Hallers-Haalboom et al., 2017).

The second aim of the present study was to elucidate the interactive verbal behaviours of parents which may promote greater balance in turn-taking in conversation. It was expected that parents' use of questions and contingent responsiveness would be positively associated with greater balance in parent-child conversational turn-taking.

Finally, the current study aimed to examine associations between parent-child conversational turn-taking and child language and cognitive abilities. In light of previous research, it was expected that greater balance in parent-child conversational turn-taking would be associated with higher child scores on standardised assessments of cognitive and language abilities. This study also sought to unpack how the components of parent-child conversation may relate to child cognitive and language skills. Again, based on previous research it was expected that parents taking longer turns would be negatively associated with child outcome measures whilst parents' use of *wh*-questions and contingent responsiveness was expected to demonstrate positive associations with child language and cognitive skills.

Children's turn-taking proficiency increases with age (Rutter & Durkin, 1987; Casillas et al., 2016) and by age two years turn-taking between parent and child is carried out with relative fluidity even in the presence of delays, irrelevant responses, and non-responding (Cekaite, 2013; Casillas et al., 2016). As mentioned previously, conversational turn-taking between parent and child within this time period may be particularly salient for later development (Gilkerson et al., 2018). This study therefore proposed to investigate the dynamics of parent-child conversational turn-taking at child age two years. Furthermore, this study observed conversational turn-taking between parent and child during structured play. Research suggests that parents are spending increasing amounts of time in structured play with their young children with a view to preparing children for school (Hirsh-Pasek et al., 2009), yet there is little research examining parental-child interaction in this context. By decomposing the construct of conversational turn-taking and investigating how specific features of both the mother-child and the father-child communicative environment at age two years are associated with turn-taking as well as child cognitive and language abilities, the

findings may provide important insights which can inform future interventions.

Method

Participants

Eighty children aged between 21–27 months (41 females; *M* = 24.06 months, *SD* = 1.39) and their biological mothers and fathers were recruited to take part in the current study. Participants were recruited through social media, flyers distributed to crèches and supermarkets, and snowballing. All participating families were White and predominantly classified as middle-class. All children included in the current study were born full-term and were typically developing. Parents were monolingual, Irish-English speaking, and residing in the family home. Mothers were aged between 25 and 46 years (*M* = 35.03, *SD* = 4.14). Fathers were aged between 23 and 55 years (*M* = 36.5, *SD* = 5.06). All mothers had completed second-level education, 77.5% had a bachelor's degree, and 35% had a postgraduate qualification. 93.8% of fathers had completed second-level education, 63.8% had a bachelor's degree, and 22.5% had a postgraduate qualification.

Procedure

The study was conducted at an Infant and Child Research Lab based in a university setting with the approval of the relevant Research Ethics Committee. Informed consent was obtained from participants prior to commencement of testing. The lab visit consisted of a developmental assessment with the child and video-recorded observations of mother-child and father-child interaction during structured play. Each child was recorded at play with their mother and father separately, thus 160 observations were recorded in total.

In the structured play condition, dyads were presented with a magnetic puzzle board (of either fish or car design) which differed between the mother-child and father-child interactions. The task firstly required the child to use a magnetic stick attached to a string (similar to a fishing-rod) to pull out ten puzzle pieces, and secondly to replace these pieces back into the correct slots once all had been removed. The task was challenging for two-year-olds and required parental input to be completed. The duration of the structured play condition was five minutes and parents were instructed to play with their children as they would at home. The order of mother-child and father-child play interactions was counterbalanced.

Interactions were video recorded using Mangold VideoSync Pro 1.5 and transcribed offline by trained research assistants using the Computerised Language Analysis (CLAN) software according to the Codes for Human Analysis of Transcripts (CHAT) conventions (MacWhinney, 2000). All speech was transcribed verbatim. These transcripts were each reviewed by a senior transcriber. Parent-child conversation

variables were extracted from the transcripts using CLAN (MacWhinney, 2000). These variables included adult and child word counts, balance in conversational turn-taking (MLT ratio), mean length of turn (MLT), and proportion of questions. Alongside video footage of the interactions, parental contingent responsiveness was also coded using these transcripts.

Information on family sociodemographic factors (*what is the highest level of education (full- or part-time) which you have completed to date?*) and child developmental status (*has your child had any longstanding illness, condition or disability or were there any complications with their birth or pregnancy?*) was collected via questionnaire. Parents and child were offered breaks during the session as needed. Participants were not given monetary compensation for taking part in the study. At the end of the visit, participants were debriefed and thanked for their time.

Measures

Conversational Turn-Taking

The index of parent-child conversational turn-taking employed by the current study was mean length of turn (MLT) ratio. The MLT ratio calculation is a measure of conversational load (MacWhinney, 2000) and is calculated as a ratio of each speakers' mean length of turn. MLT was calculated by dividing the speakers' total number of utterances by their total number of turns. An utterance was defined as a unit of speech delineated by a change in intonation, pause, or change in conversational turn (MacWhinney, 2000). A turn referred to a sequence of utterances spoken by one interlocutor. CLAN calculates turns by identifying sequences of repeated speaker ID codes at the beginning of the main line in a transcript. The end of one turn is therefore delineated by the next interlocutor commencing to speak. The ratio of child-father MLT was then calculated as an index of conversational balance such that a ratio closer to one indicated greater balance. A father and child taking equally long turns of 6 utterances each, for example, would have an MLT ratio of 1. Mother-child MLT ratio was calculated in the same manner.

A measure of adult turn counts was also included in the present analyses and was produced using the MLT command in CLAN. This quantitative measure captures the total number of turns speakers took during the five-minute interaction.

Interactive Verbal Behaviours

Mothers' and fathers' turn-taking behaviours were coded from the transcripts of the structured play interactions in CLAN and from the video recordings.

Length of Turn. Parents' length of turn was measured using the MLT command

in CLAN (MacWhinney, 2000) as described above. It is important to note that although MLT is a direct component of parent-child conversational turn-taking, it gives no indication of the child's role in the language exchange. A high MLT calculated for a father, for instance, provides no information on his child's involvement in that interaction or on that child's own MLT. Table 1 provides a sample of turn-taking from one dyad in the current study. In this example, the father produced a total of three utterances over two turns and the child produced two utterances over two turns.

Table 1. ***Example of turn-taking in father-child interaction***

Speaker	Utterance
FAT	that (i)s right.
CHI	there?
FAT	yeah.
FAT	that is a red car.
CHI	red.

Note. FAT = father; CHI = child.

Questions. Frequency lists of all parental utterances containing a question mark were calculated in CLAN using the combo +s"*?*" +t*FAT command for fathers and combo +s"*?*" +t*MOT for mothers. Consistent with CHAT transcription conventions (MacWhinney, 2000), during the transcription process, attention was paid to speaker intonation and the content and context of utterances. Questions were typically characterised by a terminal rising intonation. The number of open-ended questions (i.e., questions requiring more than yes/no response) was computed (see Table 2 for an example from the current sample) and finally proportions of total questions and open-ended questions were calculated from each parent's total number of utterances.

Table 2. ***Example of open-ended questions in father-child interaction***

Speaker	Utterance
FAT	who is that?
CHI	horse.
FAT	seahorse.
FAT	where does the seahorse go?
CHI	there.

Note. FAT = father; CHI = child.

Contingent Responsiveness. Taking each child utterance as the target utterance, parents' verbal response to the child's utterance was coded for temporal and semantic contingency.

Parents' verbal response following their child's vocalisation was first coded for its temporal contiguity. If a parental response occurred within 2 seconds of the offset of the child's vocalisation it was coded as temporally contingent (TC). Parental responses which occurred outside of the 2-second timeframe following the child's vocalisation were coded as not temporally contingent (NTC). This time frame is frequently reported in the literature on maternal verbal responsiveness (e.g., Bornstein et al., 2015; Goldstein & Schwade, 2008; McGillon et al., 2013). Parental responses that began while the child was still vocalising were considered temporally contingent. In cases where a child produced more than one utterance in succession, the timing between each child utterance was checked – if there was a gap of more than 2 seconds between two successive child utterances this was coded as NTC (i.e., no temporally contingent response from parent); if the gap between successive child utterances was less than 2 seconds no code was required. In cases where parents produced more than one utterance within the 2-second timeframe following a child vocalisation, the temporal and semantic contingency of the first utterance only was considered.

Parent responses that were coded as temporally contingent to the child's preceding vocalisation were further coded for their semantic contingency to the child's utterance using the transcripts alongside video footage in order to examine the child's current focus of attention. Parent responses that were conceptually related to their child's preceding vocalisation/focus of attention were coded as semantically contingent (SC). Parent responses that were not conceptually related to the child's vocalisation and/or served to redirect the child's focus of attention were coded as not semantically contingent (NSC).

SC parental responses were those which related to the child's current focus of attention (Roth, 1987). SC responses included parental utterances which repeated a child's vocalisation; which answered a question the child had posed; which expanded upon the child's vocalisation or activity the child was engaged in; which named the object a child was attending to or one of its components; which praised or referenced the child's current activity; and clarification requests (e.g., asking the child to repeat what they had said). In Table 1, for example, taking the child utterance "there?" the father followed the child's focus of attention and provided a semantically contingent response to the child's vocalisation, "yeah". Similarly, in Table 2, the father expanded upon the child's vocalisation "horse", saying "seahorse".

NSC responses were parental utterances which occurred within 2 seconds of the child's vocalisation which was not conceptually related to the child's utterance and

referred to something outside of the child's current focus of attention (Akhtar et al., 1991). NSC utterances included those which directed the child towards a different activity and away from their current focus of attention; where parent and child were engaged in parallel toy play; where the parent commented on their own activity or object which the parent is engaged with. The majority of NSC utterances arose when parents attempted to refocus the child's attention towards the task. In one example, a child is focussed on a particular puzzle piece, however, the father responds directing the child's attention towards the magnet in order to continue with the task:

CHI: this is my truck .
FAT: see this red bit Evan?

Temporal and semantically contingent responses to child utterances were calculated as proportions of total number of child vocalisations in mother-child and father-child interaction, respectively.

All videos were coded by the first author. Two research assistants who were blind to the study hypotheses double coded 25% of the interactions chosen at random. Cohen's Kappa statistic was used to test inter-rater reliability of the temporal contingency codes (kappa = .87), and the semantic contingency codes (kappa = .83).

Child Language and Cognitive Abilities

Child language and cognitive abilities were directly assessed by a trained research assistant using the Bayley Scales of Infant and Toddler Development-Third Edition (BSID-III). The BSID-III are widely used to assess child development and have demonstrated acceptable levels of internal consistency, test-retest reliability, and concurrent validity (Bayley, 2006). The cognitive scale assesses the child's memory, ability to manipulate objects, and knowledge of concepts such as big and small. The receptive language scale assesses child vocabulary, understanding of grammar and tenses and knowledge of prepositions. The expressive scale assesses child ability to label objects, use different tenses of verbs and use prepositions. Child scaled scores on the cognitive, receptive and expressive scales were used in the present analyses. Bayley cognitive scores were missing for one child and Bayley language scores were missing for two children. These cases were not included in the final analyses.

Results

Analytic Strategy

Data analysed in the current study were drawn from a demographic questionnaire, video-recorded mother-child and father-child play interactions, and a cognitive and language developmental assessment administered to the child during a single visit to

the lab at child age two years. Data were analysed using SPSS version 26. To address the first research question, mean-level differences in mother-child and father-child conversational balance as well as differences between mothers' and fathers' interactive verbal behaviours were analysed. Second, bivariate correlations were conducted in order to examine associations between parents' interactive verbal behaviours and parent-child conversational balance. Lastly, multiple regression analyses were conducted to investigate associations between mother-child and father-child conversational turn-taking and child cognitive and language abilities.

Comparing Father-Child and Mother-Child Conversational Turn-Taking

Descriptive statistics for parent-child turn counts, conversational balance, parents' interactive verbal behaviours, as well as quantities of parent-child speech are presented in Table 3. As parental semantic contingency was only coded from temporally contiguous responses, one measure of contingent responsiveness was used in the present analyses (i.e., the proportion of parental responses which were temporally *and* semantically contingent upon the child's vocalisations). Preliminary analyses identified a number of outliers and analyses were conducted with and without these cases. Overall, the results were not affected by the presence of these outliers and therefore these cases were retained in the final dataset.

Paired *t*-tests were conducted to compare parent-child speech variables in father-child and mother-child interaction. There was no significant difference with regards to the quantity of child speech across mother-child and father-child play interactions and no difference in the quantity of mothers' and fathers' speech, as indexed by total word counts. There was greater balance in conversational turn-taking (i.e., MLT ratio was higher) during father-child interaction compared to mother-child interaction, $t(79) = 2.12$, $p = .04$, $d = 0.24$.[1] However, mothers produced more contingently responsive utterances in response to their child's vocalisations compared to fathers, $t(79) = -2.67$, $p = .01$, $d = 0.30$, whilst fathers produced more responses which were not contingent upon the child's vocalisation, $t(79) = 2.73$, $p = .01$, $d = 0.31$. Mothers in the present sample responded to child vocalisations in both a semantically and temporally contingent manner approximately 78% of the time, whilst fathers did so on average 73% of the time. There were no significant differences between mothers and fathers in relation to mean length of turn, proportion of questions, or *wh*-questions produced during interaction. Paired *t*-tests were also run to examine any differences in mothers' and fathers' turn-taking behaviours according to child gender. No differences in parent-child turn-taking were found between boys and girls.

[1] Cohen's *d* is a measure of effect size (i.e., the size of the difference between two groups). Cohen (1988) proposed that $d = 0.2$ should be considered a small effect size, $d = 0.5$ a moderate effect size, and $d = 0.8$ a large effect size.

Table 3. ***Descriptive statistics and t-tests for parent-child conversational turn-taking behaviours during father-child and mother-child interaction***

	Father-child		Mother-child			
Measure	Mean (*SD*)	Range	Mean (*SD*)	Range	Paired Differences Skewness	*t*
PAR word tokens	411.54 (117.72)	175–698	429.33 (126.37)	192–843	.12	-.96
CHI word tokens	54.85 (37.88)	2–150	49.85 (37.92)	3–191	.15	1.42
Turn count	32.32 (14.40)	4–68	28.74 (15.61)	3–74	-.67	2.00
MLT ratio	0.38 (0.21)	0.04–.94	0.33 (0.19)	0.03–0.86	-.29	2.12*
Mean length of turn	4.32 (3.58)	1.36–25.25	5.34 (5.30)	1.68–38.67	2.42	-1.51
Questions	27.33 (11.56)	0–65.93	29.21 (11.16)	4.63–61.39	.14	-1.31
Wh-questions	7.90 (5.50)	0–25.77	7.96 (5.53)	0–30.34	-.10	-.08
Contingent responsiveness	72.78 (13.71)	33.33–96.88	78.21 (13.69)	23.08–100	-.26	-2.67**
Non-semantically contingent responses	18.02 (11.68)	0–54.55	13.41 (11.72)	0–76.92	-.01	2.73**

Note. PAR = parent; CHI = child; MLT = Mean Length of Turn.

*$p < .05$; ** $p < .01$.

Parental Interactive Verbal Behaviour and Conversational Balance

The second aim of the present study was to investigate the features of parent-child communicative exchanges which were associated with children's engagement in conversation. Tables 4 and 5 present data pertaining to the associations between parents' interactive verbal behaviours and parent-child conversational balance. These data are presented separately for mothers and fathers. As several variables were not normally distributed Spearman's correlations were conducted. Fathers' use of questions was positively associated with father-child conversational balance whilst mothers' production of *wh*-questions was positively associated with mother-child conversational balance. There were no associations between parents' contingent responsiveness and parent-child conversational balance.

Conversational Balance and Child Cognitive and Language Abilities

Tables 4 and 5 also present bivariate correlations between parent-child conversational balance and child cognitive and language abilities. The role of possible covariates including child age, parental education and parents' quantity of speech input (number of word tokens) was also considered. Mothers' level of education was slightly higher than fathers' and this difference was statistically significant, $t(73) = 4.15$, $p < .001$. Child age demonstrated a significant association with child expressive language ability as well as several features of parents' turn-taking behaviour and was therefore included as a control variable in subsequent analyses. Father-child conversational balance was positively associated with child cognitive ability and mother-child conversational balance was associated with child cognitive and expressive language abilities. Mothers' and fathers' production of *wh*-questions was positively associated with child cognitive ability and mothers' *wh*-questions were also associated with child language abilities. Mothers' MLT was negatively associated with child cognitive and expressive language abilities. Finally, mothers' non-semantically contingent responding was negatively associated with child cognitive and receptive language scores. The strength of these associations ranged from weak to medium.

To examine the contribution of parent-child conversational balance to children's cognitive and language skills, multiple regression analyses were conducted. Normal probability plots of residuals alongside scatter plots of residuals were examined prior to conducting these analyses which indicated that the assumptions of multiple regression had been satisfied. Due to its associations with multiple main variables, child age was retained as a covariate. Table 6 displays the results examining associations between parent-child conversational balance and child cognitive ability, controlling for child age.

In the first model, child cognitive ability was associated with child age. In the second model, child cognitive ability was associated with mother-child MLT ratio only, $F(3,75) = 5.39$, $p = .002$. Greater balance in mother-child conversational-turn taking was associated with greater child cognitive ability. This model explained 18% of the variance in child cognitive ability. Parents' *wh*-questions and non-semantically contingent responding were added to the third model to ascertain whether these variables contributed any additional variance to child cognitive scores. MLT could not be added to the model due to issues with multicollinearity. The addition of these variables did not significantly improve the model, (significance of F change >. 05). Examining associations between parent-child conversational balance and child receptive language, controlling for child age, produced a non-significant F-test, suggesting the model did not fit the data well. Examining associations between parent-child conversational balance and child expressive language, controlling for child age, produced a significant F-test, $F(3, 74) = 4.34$, $p = .007$, $R^2 = .15$, however none of the predictors included in the model were significant.

Table 4. ***Bivariate correlations between father-child turn-taking variables and child language and cognitive abilities***

Factor	1	2	3	4	5	6	7	8	9	10	11	12
1 CHI age	1											
2 FAT education	-.05	1										
3 FAT word tokens	.13	.16	1									
4 MLT ratio	.21	-.26*	-.38**	1								
5 FAT MLT	-.18	.29*	.36**	-.97**	1							
6 FAT questions	.21	.07	.24*	.27*	-.28*	1						
7 FAT *wh*-ques-	.22	-.09	.05	.21	-.24*	.51**	1					
8 FAT SC	.16	-.00	.27*	-.22	.15	.09	.23*	1				
9 FAT NSC	-.30**	.01	.02	.03	.01	-.07	-.20	-.79**	1			
10 Bayley Cog	.19	.06	-.01	.23*	-.21	.04	.24*	.19	-.22	1		
11 Bayley Rec	.21	.04	.13	-.02	.00	-.03	.11	.16	-.12	.56**	1	
12 Bayley Exp	.32**	.09	.15	.20	-.21	.07	.22	.11	-.18	.47**	.50**	1

Note. CHI = Child; FAT = Father; MLT = Mean length of turn; SC = Semantic contingency; NSC = Non-semantic responding; Cog = Cognitive; Rec = Receptive; Exp = Expressive.

*$p < .05$; **$p < .01$.

Table 5. ***Bivariate correlations between mother-child turn-taking variables and child language and cognitive abilities***

Factor	1	2	3	4	5	6	7	8	9	10	11	12
1 CHI age	1											
2 MOT education	-.02	1										
3 MOT word tokens	.10	-.07	1									
4 MLT ratio	.34**	-.21	-.18	1								
5 MOT MLT	-.34**	.22	.10	-.96**	1							
6 MOT questions	.21	.02	.17	.16	-.21	1						
7 MOT *wh*-questions	.35**	.08	.32**	.29**	-.33**	.57**	1					
8 MOT SC	.20	-.10	.05	-.08	.06	-.01	.14	1				
9 MOT NSC	-.08	.02	-.06	-.03	.02	.01	-.10	-.70**	1			
10 Bayley Cog	.19	.10	-.13	.44**	-.40**	.17	.33**	.08	-.23*	1		
11 Bayley Rec	.21	.05	.10	.22	-.22	.15	.37**	.13	-.28*	56**	1	
12 Bayley Exp	.32**	.08	.16	.32**	-.28*	.14	.34**	.16	-.21	.47**	.50**	1

Note. CHI = Child; MOT = Mother; MLT = Mean length of turn; SC = Semantic contingency; NSC = Non-semantic responding; Cog = Cognitive; Rec = Receptive; Exp = Expressive.

*$p < .05$; **$p < .01$.

Table 6. ***Multiple regression model predicting child cognitive ability (n=79)***

	Model 1			Model 2			Model 3		
Predictors	B	SE B	β	B	SE B	β	B	SE B	β
CHI age (in	.43	.21	.23*	.19	.21	.10	-.05	.23	-.03
MOT-CHI MLT ra-				3.71	1.72	.27*	3.73	1.72	.27*
FAT-CHI MLT ra-				2.23	1.50	.17	1.84	1.49	.14
MOT *wh*-questions							.06	.05	.13
FAT *wh*-questions							.05	.05	.10
MOT NSC							-.03	.02	-.14
FAT NSC							-.03	.02	-.17

Note. CHI = Child; MOT = Mother; FAT = Father; MLT = Mean length of turn; NSC = Non-semantic responding.

*$p < .05$.

Discussion

The current study sought to provide a detailed insight into mothers' and fathers' conversational turn-taking in interaction with their two-year-old children and investigate how interactive features of parental CDS support children's engagement in conversation. This study also aimed to elucidate any associations between father-child conversational turn-taking and child cognitive and language abilities. Fathers remain underrepresented in developmental research and the inclusion of both mothers and fathers in this study is important, as it provides a closer approximation of the early interactive environment of the developing child. To our knowledge, this is the first study to provide an in-depth examination of conversational turn-taking during father-child interaction. Overall, the results indicated that there was greater balance in conversational turn-taking in father-child interaction compared to mother-child exchanges. However, father-child turn-taking did not account for any additional variance in child cognitive ability once mother-child conversational balance was controlled for. Finally, regression analyses failed to demonstrate associations between parent-child conversational turn-taking and child receptive and expressive language skills.

The first aim of the present study was to compare father-child and mother-child conversational turn-taking as well as the interactive verbal behaviours of mothers and fathers. Although there was greater balance in father-child interaction, within turns mothers were more contingently responsive to their child's vocalisations compared to fathers. There is little research examining fathers' contingent responsiveness during toddlerhood and previous research has produced inconsistent findings. Several

studies suggest that mothers and fathers are similarly sensitive to their young child's cues (e.g., Tamis-LeMonda et al., 2004; Towe-Goodman et al., 2014) whilst others indicate that mothers' display greater contingent responsiveness compared to fathers (e.g., Flippin & Watson, 2015; Schueler & Prinz, 2013).

Hallers-Haalboom and colleagues (2017) suggested that fathers' tendency to be less contingently responsive may align with their propensity to produce more questions and directive speech during parent-child interaction compared to mothers. It is frequently cited in the literature that fathers use more questions during parent-child play compared to mothers, and in particular produce more challenging *wh*-questions (Malin et al., 2014; Rowe, Coker, & Pan, 2004). However, the present study observed no significant differences in mothers' and fathers' production of questions overall, or *wh*-questions.

This study did however find that fathers produced more responses that were not semantically contingent to their child's speech compared to mothers, although the difference was small. Directive speech was a key component of non-semantically contingent talk. It may be, as Hallers-Haalboom and colleagues (2017) proposed, that fathers were more goal-oriented than mothers and therefore were more focussed on completing the task at hand than responding contingently to their child's behaviour. Future studies examining fathers' responsiveness during free play and structured play conditions may provide further insight. Whilst it has long been contended that fathers may be more challenging communicative partners to their children compared to mothers (Gleason, 1975), the present study suggests this may be borne out in their propensity to respond non-contingently to their children's vocalisations rather than their production of *wh*-questions.

The second aim of the present study was to gain insight into the ways in which these interactive verbal behaviours support children's verbal engagement in conversation. It was expected that by posing more questions and responding contingently to children's speech initiations parents would scaffold their participation in conversation. Parents' use of questions emerged as an important feature of mothers' and fathers' CDS for engaging children in conversation. *Wh*-questions in particular may encourage children to provide longer responses. Previous research has demonstrated that two-year-olds produce more syntactically complex responses to this type of question (Rowe et al., 2017). It may be of interest, in future research, to examine in more depth the complexity and length of children's responses to different types of parental *wh*-question and yes/no questions and whether this translates to children taking longer turns. Parents' contingent responsiveness was not associated with conversational balance. Perhaps, as Locke (1996) suggested, this feature of caregiver-child communication may be less important for engaging children of the current age group in back-and-forth exchanges compared to asking questions. This may also explain the lack of associations between parental responsiveness and child language and cognitive

abilities.

It is unclear from the present results how differences in mothers' and fathers' interactive behaviours were associated with differences in mother-child and father-child conversational balance. There was greater conversational balance in father-child play but there were no differences in mothers' and fathers' use of questions. Although the difference was not statistically significant, fathers' mean length of turn was shorter than mothers' mean length of turn. As previously mentioned, when caregivers decrease the length of turns they take, children's verbal participation in communicative exchanges tends to increase (Brassart & Schelstraete, 2015; Girolametto, 1988). It is also possible that a feature of turn-taking not considered in this study may account for the present findings.

Pausing, for instance, is an important unit of turn-taking which serves as a cue for speaker transitions (Schlangen, 2006). Sufficient pausing following a parental utterance ensures the child has enough time to plan and initiate their response and facilitates children's participation in conversation. More in-depth analysis of pauses between consecutive parental utterances within turns may elucidate whether parents were providing temporal space for their children to respond and whether or not children were availing of these opportunities to participate in conversation. Perhaps fathers in the current sample provided more cues regarding speaker transition through pausing which encouraged child engagement in conversation. Furthermore, the timing of parents' responses to their child's vocalisation in the current study were coded as either occurring within two seconds or not. If more detailed examination regarding the timing of these responses in milliseconds was carried out, perhaps it would emerge that fathers' timing provided more temporal space for the child to take multiple utterances per turn, thus facilitating greater balance in conversation. Future research may also benefit from examining the role of prosody, gesture and gaze as important elements of conversational turn-taking (e.g., Kuchirko et al., 2017; Rohlfing et al., 2020; Rutter & Durkin, 1987). Instances where parents may have provided prosodic or visual cues to mark turn boundaries and children did not take a subsequent turn may not be captured by the present coding scheme.

The final aim of this study was to examine concurrent associations between child language and cognitive abilities and parent-child conversational turn-taking. Whilst mother-child and father-child balance were separately correlated with child cognitive scores, regression analyses indicated that considered jointly, mother-child conversational balance was the only variable significantly associated with child cognitive ability. In other words, father-child conversational balance did not explain any unique variance in child cognitive abilities above and beyond mother-child conversational balance. Similarly, although mothers' and fathers' *wh*-questions were positively correlated with child language and cognitive competencies, these variables did not contribute any additional variance in child cognitive ability.

In the present study, the only difference observed between mothers' and fathers' interactive behaviours was that mothers were more contingent. As contingency was not associated with cognitive abilities, it may be that it interacts with a linguistic feature of mothers', and not fathers', CDS to support children's cognitive development. It may be of interest to future research to examine linguistic features of mothers' and fathers' CDS such as vocabulary diversity and language complexity and how these interact with the interactive features of parents' CDS to influence child development. It may also be important to consider whether the MLT ratio measure employed in the current study favours parents' use of shorter utterances which could lead to simpler speech on the part of the parent. Future studies may address this concern by examining associations between parents' language complexity and balance in parent-child conversational turn-taking.

It is also possible that longitudinal associations may emerge between father-child turn-taking and child cognitive and language development. Previous research has suggested that certain aspects of fathers' parenting may exert specific influences on child development at certain points in time (Towe-Goodman et al., 2014). It is conceivable that over a longer period of time, the effects of father-child conversational turn-taking on child cognitive and language abilities would be elucidated. It is also possible that the current study was underpowered to demonstrate associations between father-child conversational turn-taking and child abilities after controlling for mother-child turn taking. Nonetheless, participation in conversation likely relies on several cognitive skills such as attention and executive function (Casillas et al., 2016) and the present results indicate that the contribution of mother-child conversational turn-taking to child cognitive development is important, despite being less balanced compared to father-child turn-taking.

Strengths and Limitations

This study adds to our knowledge on the dynamics of parent-child conversation and is one of few studies to examine turn-taking within the father-child dyad. The inclusion of both mothers and fathers in the current study permitted a closer approximation of the children's early interactive environment compared to previous research, which has primarily focussed on mother-child exchanges. The use of observational methods to capture naturalistic interactions between parents and children is considered gold standard in the field of fathering research (Cabrera & Volling, 2019). The lab setting also allowed for stimuli and environmental factors to be controlled for across all participants, facilitating comparability across the present sample (De Barbaro et al., 2013). Direct assessment of child cognitive and language skills was another strength of the research as this provided an objective measure of child abilities. Parent-report measures of child capabilities or behaviour may be subject to social desirability and recall bias (Baumeister et al., 2009; Chorney et al., 2014). Finally, the

present design mitigated several limitations of the LENA device mentioned previously.

The cross-sectional design of the current study, however, makes it difficult to tease apart the direction of influences between parent and child factors under consideration. Longitudinal analyses which control for children's baseline abilities may elucidate the direction of the associations between conversational turn-taking and child development over time. For instance, parents may take longer turns when children have lower language abilities. Longitudinal analyses would also allow us to examine the bidirectional associations between turn-taking and child developmental capacities. It is also important to consider how the brief play interactions measured in the lab environment represent the daily experiences of parents and children. Despite advantages of studying behaviour in a laboratory setting, as discussed above, behaviours measured in this setting may have lower ecological validity than observations taken in the home.

The variables included in the present analyses accounted for a small percentage of the variance in child cognitive ability and, as previously mentioned, factors which were not included in the present study likely have important implications for children's development. Data on child birth order, for example, were not compiled. Whilst some research suggests that parent-child dynamics and development may be impacted by child birth order (e.g., Bornstein et al., 2019; Lehmann et al., 2016), other research has not observed an effect of birth order on mothers' and fathers' behaviours during parent-child interaction (e.g., Hallers-Haalboom et al., 2017).

The homogeneity of the sample, which comprised White, highly educated, married parents, may limit the generalisability of the current findings. There may have been limited variability in mothers' and fathers' conversational turn-taking and interactive verbal behaviours in the present sample compared to more diverse populations. This is important to acknowledge given established associations between socioeconomic status and CDS (Schwab & Lew-Williams, 2016). Whilst maternal education is perhaps the strongest predictor of child language development (McNally et al., 2019), there was little variation in this domain among the current sample in order to control for such effects.

Whilst mothers in the current sample were slightly more educated than fathers, fathers' education, and not mothers', was significantly associated with fathers' conversational balance and mean length of turn. On the other hand, mothers' mean length of turn was negatively associated with child age. It is possible that mothers are taking shorter turns with slightly older children to signal greater responsibility for them to engage in the back-and-forth exchange. Future research with a more diverse sample may allow for the associations between sociodemographic factors and parent-child conversational behaviours to be teased apart more clearly.

Finally, it is important to acknowledge cultural assumptions regarding developmental milestones and processes of development (Kuchirko & Nayfeld, 2020). For instance, there are communities outside of the Western world where CDS is relatively rare (e.g., Casillas et al., 2020) and different cultures may have distinct expectations for children's verbal participation in interaction (Girolametto et al., 2002). Participants in the present collection of studies were also homogeneous in relation to family composition. Families comprised two-parent households consisting of a biological resident father and mother. It may therefore be important to consider family structure when generalising the present findings and when making comparisons across future replications.

Conclusion

In order to attain a more comprehensive account of the developing child's early environment it is crucial to consider the multiple contexts within which a child develops. Research on both mother-child and father-child interaction provides an important insight into the early interactive experiences of children and how this shapes their development. Results from this study provide a deeper understanding of the processes by which fathers and mothers interact with their children during conversation and indicate that taking shorter turns and using questions is associated with greater balance in conversational turn-taking between parent and child. The results also added to the small body of research on the role of pragmatics in child cognitive development. Promoting "serve and return" interactions between parents and children may have significant implications for children's development and equip children with the skills needed for future success. Future research with a larger, more socioeconomically diverse sample is however needed to test longitudinal associations between father-child conversational turn-taking and child development.

References

Akhtar, N., Dunham, F., & Dunham, P. J. (1991). Directive interactions and early vocabulary development: The role of joint attentional focus. *Journal of Child Language, 18*(1), 41–49. https://doi.org/10.1017/s0305000900013283

Baker, C. E., & Vernon-Feagans, L. (2015). Fathers' language input during shared book activities: Links to children's kindergarten achievement. *Journal of Applied Developmental Psychology, 36*, 53–59. https://doi.org/10.1016/j.appdev.2014.11.009

Bateson, M. C. (1979). Parent-infant exchanges: The epigenesis of conversational interaction: A personal account of research and development. In M. Bullowa (Ed.), *Before speech: The beginnings of human communication* (pp. 63–77). Cambridge University Press.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2009). Psychology as the science of self-reports and finger movements: Whatever happened to actual behaviour? Perspectives on Psychological Science, 2(4), 396-403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development - Third Edition*. Harcourt Assessment, Inc.

Bornstein, M. H., Putnick, D. L., Cote, L. R., Haynes, O. M., & Suwalsky, J. T. D. (2015). Mother-infant contingent vocalizations in 11 countries. *Psychological Science, 26*(8), 1272–1284. https://doi.org/10.1177/0956797615586796

Bornstein, M. H., Putnick, D. L., & Suwalsky, J. T. D. (2019). Mother–infant interactions with firstborns and secondborns: A within-family study of European Americans. *Infant Behavior and Development, 55*, 100–111. https://doi.org/10.1016/j.infbeh.2019.03.009

Bornstein, M. H., Tamis-LeMonda, C. S., & Haynes, O. M. (1999). First words in the second year: Continuity, stability, and models of concurrent and predictive correspondence in vocabulary and verbal responsiveness across age and context. *Infant Behavior & Development, 22*(1), 65–85. https://doi.org/10.1016/S0163-6383(99)80006-X

Bowlby, J. (1969). *Attachment and loss* (Vol.1). Basic Books.

Brassart, E., & Schelstraete, M.-A. (2015). Simplifying parental language or increasing verbal responsiveness, what is the most efficient way to enhance pre-schoolers' verbal interactions? *Journal of Education and Training Studies, 3*(3). https://doi.org/10.11114/jets.v3i3.709

Bruner, J. S. (1981). The social context of language acquisition. *Language & Communication, 1*(2/3), 155–78. https://doi.org/10.1016/0271-5309(81)90010-0

Bruner, J. S. (1983). *Child's talk: Learning to use language*. Oxford University Press.

Cabrera, N. J., Fitzgerald, H. E., Bradley, R. H., & Roggman, L. (2014). The ecology of father-child relationships: An expanded model. *Journal of Family Theory & Review, 6*(4), 336-354. https://doi.org/10.1111/jftr.12054

Cabrera, N. J., & Volling, B. L. (2019). Moving research on fathering and children's development forward: Priorities and recommendations for the future. In B. L.

Volling & N. J. Cabrera (Eds.), *Advancing research and measurement on fathering and children's development. Monographs of the Society for Research in Child Development, 84*(1), 107–117. https://doi.org/10.1002/mono.12404

Casillas, M., Bobb, S. C., & Clark, E. V. (2016). Turn-taking, timing, and planning in early language acquisition. *Journal of Child Language, 43*(6), 1310–1337. https://doi.org/10.1017/s0305000915000689

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development, 91*(5), 1819-1835. https://doi.org/10.1111/cdev.13349

Cekaite, A. (2013). Child pragmatic development. In *The Encyclopaedia of Applied Linguistics* (pp. 1–7). Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0127

Chorney, J. M., McMurtry, C. M., Chambers, C. T., & Bakeman, R. (2014). Developing and modifying behavioral coding schemes in paediatric psychology: A practical guide. *Journal of Pediatric Psychology, 40*(1), 154–164. https://doi.org/10.1093/jpepsy/jsu099

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Conica, M., Nixon, E., & Quigley, J. (2020). Fathers' but not mothers' repetition of children's utterances at age two is associated with child vocabulary at age four. *Journal of Experimental Child Psychology*, 104738. https://doi.org/10.1016/j.jecp.2019.104738

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J. & Bergelson, E. (2020). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01393-5

De Barbaro, K., Johnson, C. M., Forster, D., & Deak, G. O. (2013). Methodological considerations for investigating the microdynamics of social interaction development. *IEEE Transactions on Autonomous Mental Development, 5*(3), 258–270. https://doi.org/10.1109/tamd.2013.2276611

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*. https://doi.org/10.1111/cdev.13511

Ferjan Ramírez, N., Hippe, D. S., & Kuhl, P. K. (2021). Comparing automatic and manual measures of parent–infant conversational turns: A word of caution. *Child Development*. https://doi.org/10.1111/cdev.13495

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development, 60*(6), 1497-1510. https://doi.org/10.2307/1130938

Flippin, M., & Watson, L. R. (2015). Fathers' and mothers' verbal responsiveness and the language skills of young children with autism spectrum disorder. *American Journal of Speech-Language Pathology, 24*(3), 400–410. https://doi.org/10.1044/2015_AJSLP-13-0138

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J.H.L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology, 26*(2), 248–265. https://doi.org/10.1044/2016_ajslp-15-0169

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*(4), e20174276. https://doi.org/10.1542/peds.2017-4276

Girolametto, L. (1988). Improving the social-conversational skills of developmentally delayed children: An intervention study. *Journal of Speech & Hearing Disorders, 53*(2), 156-167. https://doi.org/10.1044/jshd.5302.156

Girolametto, L., Bonifacio, S., Visini, C., Weitzman, E., Zocconi, E., & Pearce, P. S. (2002). Mother-child interactions in Canada and Italy: Linguistic responsiveness to late-talking toddlers. *International Journal of Language & Communication Disorders, 37*(2), 153–171. https://doi.org/10.1080/13682820110116794

Gleason, J. B. (1975). Fathers and other strangers: Men's speech to young children. In I. D. P. Dato (Ed.), *Developmental psycholinguistics: Theory and applications* (pp. 289-297). Georgetown University Press.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Physiological Science, 19*(5), 515–23. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Gómez, E., & Strasser, K. (2021). Language and socioemotional development in early childhood: The role of conversational turns. *Developmental Science*, e13109. https://doi.org/10.1111/desc.13109

Hallers-Haalboom, E. T., Groeneveld, M. G., van Berkel, S. R., Endendijk, J. J., van der Pol, L. D., Linting, M., Bakermans-Kranenburg, M. J., & Mesman, J. (2017). Mothers' and fathers' sensitivity with their two children: A longitudinal study from infancy to early childhood. *Developmental Psychology, 53*(5), 860–872. https://doi.org/10.1037/dev0000293

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P.K.S., & Suma, K. (2015). The contribution of early communication quality to low-income children's language success. *Psychological Science, 26*, 1071–1089. https://doi.org/10.1177/0956797615581493

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., & Singer, D. G. (2009). *A mandate for playful learning in the preschool: Presenting the evidence*. Oxford University Press.

Kopp, C. B. (1982). Antecedents of self-regulation: A developmental perspective. *Developmental Psychology, 18*(2), 199–214. https://doi.org/10.1037/0012-1649.18.2.199

Kuchirko, Y., & Nayfeld, I. (2020). Language gap: Cultural assumptions and ideologies. In C. Huertas-Abril & M. Gómez-Parra (Eds.), *International approaches to bridging the language gap* (pp. 32-53). IGI Global https://doi.org/10.4018/978-1-7998-1219-7.ch003

Kuchirko, Y., Tafuro, L., & Tamis LeMonda, C. S. (2017). Becoming a communicative partner: Infant contingent responsiveness to maternal language and gestures. *Infancy, 23*(4), 558–576. https://doi.org/10.1111/infa.12222

Lamb, M. E., & Lewis, C. (2010). The development and significance of father–child relationships in two-parent families. In M. E. Lamb (Ed.). *The role of the father in child development* (5th ed., pp. 94-153). John Wiley & Sons.

Landry, S. H., Smith, K. E., Swank, P. R., & Miller-Loncar, C. L. (2000). Early maternal and child influences on children's later independent cognitive and social functioning. *Child Development, 71*(2), 358-375. https://doi.org/10.1111/1467-8624.00150

Leech, K. A., Salo, V. C., Rowe, M. L., & Cabrera, N. J. (2013). Father input and child vocabulary development: The importance of *wh*-questions and clarification requests. *Seminars in Speech and Language, 34*(4), 249–259. https://doi.org/10.1055/s-0033-1353445

Lehmann, J.-Y. K., Nuevo-Chiquero, A., & Vidal-Fernandez, M. (2016). The early origins of birth order differences in children's outcomes and parental behavior. *Journal of Human Resources, 53*(1), 123–156. https://doi.org/10.3368/jhr.53.1.0816-8177

Lloyd, J., Lieven, E., & Arnold, P. (2001). Oral conversations between hearing-impaired children and their normally hearing peers and teachers. *First Language, 21*(61), 83–107. https://doi.org/10.1177/014272370102106104

Locke, J. L. (1996). Why do infants begin to talk? Language as an unintended consequence. *Journal of Child Language, 23*(2), 251–268. https://doi.org/10.1017/s0305000900008783

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Malin, J. L., Cabrera, N. J., & Rowe, M. L. (2014). Low-income minority mothers' and fathers' reading and children's interest: Longitudinal contributions to children's receptive vocabulary skills. *Early Childhood Research Quarterly, 29*(4), 425–432. https://doi.org/10.1016/j.ecresq.2014.04.010

Masur, E. F., Flynn, V., & Eichorst, D. L. (2005). Maternal responsive and directive behaviours and utterances as predictors of children's lexical development. *Journal of Child Language, 32*(1), 63–91. https://doi.org/10.1017/s0305000904006634

McDonnell, S. A., Friel-Patti, S., & Rosenthal Rollins, P. (2003). Patterns of change in maternal–child discourse behaviors across repeated storybook readings. *Applied Psycholinguistics, 24*(3), 323–341. https://doi.org/10.1017/s0142716403000171

McGillion, M. L., Herbert, J. S., Pine, J. M., Keren-Portnoy, T., Vihman, M. M., & Matthews, D. E. (2013). Supporting early vocabulary development: What sort of responsiveness matters? *IEEE Transactions on Autonomous Mental Development, 5*(3), 240–248. https://doi.org/10.1109/tamd.2013.2275949

McNally, S., McCrory, C., Quigley, J., & Murray, A. (2019). Decomposing the social gradient in children's vocabulary skills at 3 years of age: A mediation analysis using data from a large representative cohort study. *Infant Behavior and Development, 57*, 101326. https://doi.org/10.1016/j.infbeh.2019.04.008

Pancsofar, N., & Vernon-Feagans, L. (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology, 27*(6), 571–587. https://doi.org/10.1016/j.appdev.2006.08.003

Pancsofar, N., & Vernon-Feagans, L. (2010). Fathers' early contributions to children's language development in families from low-income rural communities. *Early Childhood Research Quarterly, 25*(4), 450–463. https://doi.org/10.1016/j.ecresq.2010.02.001

Rohlfing, K. J., Leonardi, G., Nomikou, I., Raczaszek-Leonardi, J., & Hullermeier, E. (2020). Multimodal turn-taking: Motivations, methodological challenges, and novel approaches. *IEEE Transactions on Cognitive and Developmental Systems, 12*(2), 260–271. https://doi.org/10.1109/tcds.2019.2892991

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science, 29*(5), 700–710. https://doi.org/10.1177/0956797617742725

Roth, P. L. (1987). Temporal characteristics of maternal verbal styles. In Nelson, K. E., van Kleeck, A. (Eds.), *Children's language* (Vol. 6, pp. 137–159). Lawrence Erlbaum Associates.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Rowe, M. L., Coker, D., & Pan, B. A. (2004). A comparison of fathers' and mothers' talk to toddlers in low-income families. *Social Development, 13*(2), 278-291. https://doi.org/10.1111/j.1467-9507.2004.000267.x

Rowe, M. L., Leech, K. A., & Cabrera, N. (2017). Going beyond input quantity: "*Wh*"-questions matter for toddlers' language and cognitive development. *Cognitive Science, 41*, 162-179. https://doi.org/10.1111/cogs.12349

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language, 47*(1), 5–21. https://doi.org/10.1017/s0305000919000655

Rutter, D. R., & Durkin, K. (1987). Turn-taking in mother–infant interaction: An examination of vocalizations and gaze. *Developmental Psychology, 23*(1), 54–61. https://doi.org/10.1037/0012-1649.23.1.54

Sameroff, A. J. (2009). *The transactional model of development: How children and contexts shape each other*. American Psychological Association.

Schlangen, D., (2006). From reaction to prediction: Experiments with computational models of turn-taking. In *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, Pittsburgh, USA. https://pub.uni-bielefeld.de/record/1992227

Schueler, C. M., & Prinz, R. J. (2013). The role of caregiver contingent responsiveness in promoting compliance in young children. *Child Psychiatry & Human Development, 44*(3), 370–381. https://doi.org/10.1007/s10578-012-0331-0

Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(4), 264–275. https://doi.org/10.1002/wcs.1393

Schwab, J. F., Rowe, M. L., Cabrera, N., & Lew-Williams, C. (2018). Fathers' repetition of words is coupled with children's vocabularies. *Journal of Experimental Child Psychology, 166*, 437–450. https://doi.org/10.1016/j.jecp.2017.09.012

Snow, C. E. (1977). Mothers' speech research: from input to interaction In Snow, C. E. & Ferguson, C. A. (Eds.), *Talking to children: language input and acquisition* (pp. 31–49). Cambridge University Press.

Tamis-LeMonda, C. S., & Bornstein, M. H. (2002). Maternal responsiveness and early language acquisition. In R. V. Kail & Reese, H. W. (Eds.) *Advances in Child Development and Behaviour, 29*, 89-127. https://doi.org/10.1016/S0065-2407(02)80052-0

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development, 72*(3), 748-767. https://doi.org/10.1111/1467-8624.00313

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science, 23*, 121–126. https://doi.org/10.1177/0963721414522813

Tamis-LeMonda, C. S., Shannon, J. D., Cabrera, N. J., & Lamb, M. E. (2004). Fathers and mothers at play with their 2- and 3-year-olds: Contributions to language and cognitive development. *Child Development, 75*(6), 1806–1820. https://doi.org/10.1111/j.1467-8624.2004.00818.x

Towe-Goodman, N. R., Willoughby, M., Blair, C., Gustafsson, H. C., Mills-Koonce, W. R., & Cox, M. J. (2014). Fathers' sensitive parenting and the development of early executive functioning. *Journal of Family Psychology, 28*(6), 867–876. https://doi.org/10.1037/a0038128

Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry, 42*(1), 3–48. https://doi.org/10.1017/s0021963001006552

Vaughan, J., Wigglesworth, G., Loakes, D., Disbray, S., & Moses, K. (2015). Child-caregiver interaction in two remote Indigenous Australian communities. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00514

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 43*(2), 230–246. https://doi.org/10.1037/h0084224

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics, 124*(1), 342–349. https://doi.org/10.1542/peds.2008-2267

Data, Code and Materials Availability Statement

The raw data, analysis syntax, and transcripts used in the current study are available on the Open Science Framework at https://osf.io/h8czg/.

Ethics Statement

Ethical approval for the present study was obtained from the ethics committee of Trinity College Dublin. All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

LK was involved in the conceptualisation of the research, data collection, data curation, coding and analysis, and wrote the first draft of the manuscript. LN was involved in the conceptualisation of the research, methodology, provision of resources, data curation and analysis, review and editing of the manuscript, and supervision. JQ was involved in the conceptualisation of the research, methodology, provision of resources, data curation and analysis, review and editing of the manuscript, and supervision. All authors gave final approval of the version of the manuscript to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

The authors wish to thank all the families who took part in the research as well as the team at the Infant and Child Research Lab for their support.

License

Non-word repetition in children learning Yélî Dnye

Alejandrina Cristia
Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University

Marisa Casillas
Max Planck Institute for Psycholinguistics
University of Chicago

Abstract: In non-word repetition (NWR) studies, participants are presented auditorily with an item that is phonologically legal but lexically meaningless in their language, and asked to repeat this item as closely as possible. NWR scores are thought to reflect some aspects of phonological development, saliently a perception-production loop supporting flexible production patterns. In this study, we report on NWR results among children (N = 40, aged 3–10 years) learning Yélî Dnye, an isolate language spoken on Rossel Island in Papua New Guinea. Results make three contributions that are specific, and a fourth that is general. First, we found that non-word items containing typologically frequent sounds are repeated without changes more often than non-words containing typologically rare sounds, above and beyond any within-language frequency effects. Second, we documented rather weak effects of item length. Third, we found that NWR scores correlate strongly with age, whereas they are only weakly correlated with child sex, maternal education, and birth order. Fourth, we weave our results with those of others to serve the general goal of reflecting on how NWR scores can be compared across participants, studies, languages, and populations, and the extent to which they shed light on the factors universally structuring variation in phonological development at a global and individual level.

Keywords: phonology, non-word repetition, Papuan, non-industrial, non-urban, comparative, typology, markedness, literacy

Corresponding author(s): Alejandrina Cristia, 29, rue d'Ulm, 75005 Paris, France. E-mail: alecristia@gmail.com

ORCID ID(s): http://orcid.org/0000-0003-2979-4556 ; https://orcid.org/0000-0001-5417-0505

Citation: Cristia, A. & Casillas, M. (2022). Non-word repetition in children learning Yélî Dnye. *Language Development Research, 2*(1), 69—104. http://doi.org/10.34842/zr2q-1x28

Introduction

Children's perception and production of phonetic and phonological units continues developing well beyond the first year of life, even extending into middle childhood (e.g., Hazan & Barrett, 2000; Rumsey, 2017). Some of the evidence for later phonological development comes from non-word repetition (NWR) tasks. In the present study, we use NWR to investigate the phonological development of children learning Yélî Dnye, an isolate language spoken in Papua New Guinea (PNG), which has a large and unusually dense phonological inventory. This allows us to contribute data at the intersection of language typology, language acquisition, and individual variation, as presented in more detail below.

Defining NWR

In a basic NWR task, the participant listens to a production of a word-like form, such as /bilik/, and then repeats back what they heard without changing any phonological feature that is contrastive in the language. For instance, in English, a response of [bilig] or [pilik] would be scored as incorrect; a response [biːlik], where the vowel is lengthened without change of quality would be scored as correct, because English does not have contrastive vowel length.

NWR has been used to seek answers to a variety of theoretical questions, including what the links between phonology, working memory, and the lexicon are (Bowey, 2001), and how extensively phonological constraints found in the lexicon affect online production (Gallagher, 2014). NWR is also frequently used in applied contexts, notably as a diagnostic tool for language delays and disorders (Chiat, 2015; Estes, Evans, & Else-Quest, 2007). Since non-words can be generated in any language, it has attracted the attention of researchers working in multilingual and linguistically diverse environments, particularly in Europe in the context of diagnosing language impairments among bilingual children (Armon-Lotem, Jong, & Meir, 2015; Chiat, 2015; COST Action, 2009; Meir, Walters, & Armon-Lotem, 2016). NWR tasks probably tap into many skills (for relevant discussion see Coady & Evans, 2008; Santos, Frau, Labrevoit, & Zebib, 2020). Non-words can be designed to try to isolate certain skills more narrowly; for instance, one can choose non-words that contain real morphemes in order to load more on prior language experience, or non-words that are shorter to avoid loading on working memory (see a discussion in Chiat, 2015). Broadly, however, NWR scores will necessarily reflect to a certain extent phonological knowledge (to perceive the item precisely despite not having heard it before) as well as online phonological working memory (to encode the item in the interval between hearing it and saying it back) and flexible production patterns (to produce the item precisely despite not having pronounced it before).

The Present Work

We aimed to contribute to four areas of research. We motivate each in turn.

NWR and Typology

The first research area is at the intersection of typology and phonological development. There has been an interest in adapting NWR to different languages, in part for applied purposes. In a review of NWR as a potential task to diagnose language impairments among bilingual children in Europe, Chiat (2015) discusses the impossibility of creating language-universal non-word items: Languages vary in their phonological inventory, sound sequencing (phonotactics), syllable structure, and word-level prosody. As a result, any one item created will be relatively easier if it more closely resembles real words in a language, making it difficult to balance difficulty when comparing children learning different languages. This previous literature also suggests some dimensions of difficulty—an issue to which we return in the next subsection.

Although this cross-linguistic literature is rich, the potential difficulty associated with specific phonetic targets composing the non-words has received relatively little attention. For example, Chiat (2015) discusses segmental complexity as a function of whether there are consonant clusters – which is arguably a factor reflecting phonotactics and syllable structure.

In the present study, we thought it was relevant to represent the rich phonological inventory found in Yélî Dnye by including a variety of phonetic targets. Some of them are cross-linguistically rare, in that they are less common across languages than other sounds or phonetic targets. Phonologists, phoneticians, and psycholinguists have discussed the extent to which cross-linguistic frequency may reflect ease of processing and acquisition via diachronic language change. These works focus largely on phonotactics (Moreton & Pater, 2012), perceptual parsing of the (ambiguous) linguistic signal (Beddor, 2009; Ohala, 1981), and individual differences in processing styles (Bermúdez-Otero, 2015); which are small effects that may nonetheless cumulatively drive language change via phonologization (see Yu, 2021 for a recent review). Thus, the correlation between typological frequency and ease of acquisition is typically assumed to emerge from one or more of the following causal paths:

1. Sounds (and sound sequences) that are harder to perceive tend to be misperceived and thus lost diachronically
2. Sounds (and sound sequences) that are harder to pronounce tend to be mispronounced and thus lost diachronically
3. Sound sequences that are harder to hold in memory tend to be mispronounced and thus lost diachronically

Since NWR can tap into perception, production, and working memory, we predicted that variation in NWR across items will correlate with the cross-linguistic frequency of the phones composing those items.

Length Effects on NWR

The second research area we contribute data to is research looking at the impact of

word length on NWR repetition within specific languages. Some work documents much lower NWR scores for longer, compared to shorter, items (e.g., among Cantonese-learning children, Stokes, Wong, Fletcher, & Leonard, 2006), whereas differences are negligible in other studies (e.g., among Italian learners, Piazzalunga, Previtali, Pozzoli, Scarponi, & Schindler, 2019).

It is possible that differences are due to language-specific characteristics, including the most common length of words in the lexicon and/or in child-experienced speech in that culture—a hypothesis discussed for instance in Chiat (2015) (pp. 7-8; see also p. 5). In broad terms, one may expect languages with a lexicon that is heavily biased towards monosyllables to show greater length effects than languages where words tend to be longer. A non-systematic meta-analysis does not provide overwhelming support for this hypothesis (Cristia & Casillas, 2021, SM1).

Nonetheless, given the paucity of research looking at this question, and the diversity of current results, we did not approach this issue within a hypothesis-testing framework but sought instead to provide additional data on the question, which may be reused in future meta- or mega-analyses.

Individual Variation Correlations with NWR

The third research area we contribute data to relates to the possibility that children differ from each other in NWR scores in systematic ways. Although the ideal systematic review is missing, a recent paper comes close with a rather extensive review of the literature looking at correlations between NWR scores and a variety of child-level variables, including familial socio-economic status, child vocabulary, and, among multilingual children, levels of exposure to the language on which the non-words are based (Farabolini, Rinaldi, Caselli, & Cristia, 2021). In a nutshell, most evidence is mixed, suggesting that individual variation effects may be small, and more data is needed to estimate their true size. For this reason, we descriptively report association strength between NWR scores and child age, sex, birth order, and maternal education.

Our focus on age stems from previous work, where performance increases with child age (Farmani et al., 2018; Kalnak, Peyrard-Janvid, Forssberg, & Sahlén, 2014; Vance, Stackhouse, & Wells, 2005). Although past research has not investigated potential correlations with birth order on NWR, there is a sizable literature on these correlations in other language tasks (e.g., Havron et al., 2019), and therefore we report on these too. Common explanations for advantages for first- over later-born children include differential allocation of familial resources, particularly parental behaviors of cognitive stimulation (Lehmann, Nuevo-Chiquero, & Vidal-Fernandez, 2018). Regarding child sex, no significant correlation has been found in previous NWR research (Chiat & Roy, 2007), and in other language tasks evidence is mixed. Finally, prior research using NWR varies on whether significant differences as a function of maternal education are reported. For instance, no significant differences were found in some studies (Balladares, Marshall, & Griffiths, 2016; Farmani et al., 2018; Kalnak et al., 2014;

Meir & Armon-Lotem, 2017); whereas significant differences were reported in others (Santos et al., 2020; Tuller et al., 2018). In other lines of work, maternal education often correlates with child language outcomes, including vocabulary reports (Frank, Braginsky, Yurovsky, & Marchman, 2017) and word comprehension studies (Scaff, 2019). The causal pathways explaining this correlation are complex, but one explanation that is often discussed involves more educated mothers talking more to their children (see discussion in Cristia, Farabolini, Scaff, Havron, & Stieglitz, 2020).

NWR as a Function of Language and Culture

The fourth research goal we pursued is to use NWR with non-Western, non-urban populations, speaking a language with a moderate to large phonological inventory (see Maddieson, 2005 for a broad classification of languages based on inventory size). Indeed, NWR has seldom been used outside of urban settings in Europe and North America (Cristia et al., 2020; with exceptions including Gallagher, 2014). To our knowledge, it has never been used with speakers of languages having large phonological inventories (e.g., more than 34 consonants and 7 vowel qualities; Maddieson, 2013b, 2013a).

There are no theoretical reasons to presume that the technique will not generalize to these new conditions. That said, Cristia et al. (2020) recently reported relatively lower NWR scores among the Tsimane', a non-Western rural population, interpreting these findings as consistent with the hypothesis that lower levels of infant-directed speech and/or low prevalence of literacy in a population could lead to population-level differences in NWR scores.

In view of these results, it is important to bear in mind that NWR is a task developed in countries where literacy is widespread, and it is considered an excellent predictor of reading (for instance, better than rhyme awareness, e.g., Gathercole, Willis, & Baddeley, 1991). Therefore, it may not be a general index of phonological development, but instead reflect certain non-universal language skills. Indeed, Cristia et al. (2020) present their task as being a good index of the development of "short-hand-like" representations specifically, which could thus miss, for example, more holistic phonological and phonetic representations. We return to the question of what was measured here in the Discussion.

Aside from Cristia et al. (2020)'s hypotheses just mentioned, we have found little discussion of linguistic differences (i.e., potential differences in NWR as a function of which specific language children are learning, and/or its typology) or cultural differences (i.e., potential differences in NWR as a function of other differences across human populations).[1]

[1] Please note that the linguistic and cultural differences discussed here are different from the differences discussed in the extensive literature on NWR by bilingual participants. In that literature, authors are concerned with individual variation in exposure to one (as opposed to other) languages among multilingual children, as variation

Regarding potential language differences, we note that previous studies composed items by varying syllable structure and word length, while preferring relatively simple and universal phones (notably relying on point vowels, simple plosives, and fricatives that are prevalent across languages, like /s/). It would be interesting for future researchers to consider straying from the literature by varying other dimensions that are relevant to the language under study. For instance, for Yélî Dnye, it is relevant to vary phonological complexity of the individual sounds because of its large inventory.

Yélî Dnye phonology and community.

Before going into the details of our study design, we first give an overview of Yélî Dnye phonology as well as a brief ethnographic review of the developmental environment on Rossel Island. As discussed above, NWR has been almost exclusively used in urban, industrialized populations, so we provide this additional ethnographic information to contextualize the adaptations we have made in running the task and collecting the data, compared to what is typical in commonly studied sites. Rossel Island lies 250 nautical miles off the coast of mainland PNG and is surrounded by a barrier reef. As a result, transport to and from the island is both infrequent and irregular. International phone calls and digital exchanges that require significant data transfer are typically not an option. Data collection is therefore typically limited to the duration of the researchers' on-island visits.

Yélî Dnye Phonology

Yélî Dnye is an isolate language (presumed Papuan) spoken by approximately 7,000 people residing on Rossel Island, an island found at the far end of the Louisiade Archipelago in Milne Bay Province, Papua New Guinea. The Yélî sound system, much like its baroque grammatical system (Levinson, 2021), is unlike any other in the region. In total, Yélî Dnye uses 90 distinctive segments (not including an additional three rarely used consonants), far outstripping the phoneme inventory size of other documented Papuan languages (Foley, 1986; Levinson, 2021; Maddieson & Levinson,

in relative language experiences could mask potential effects of language impairment. To try to measure language abilities above and beyond relative levels of experience with a given language, authors have tried to build non-words that tap language-dependent or language-independent knowledge. For instance, Tuller et al. (2018) employed a set of non-words judged to be language independent and two others that were more aligned with either French or German. The intuition is that NWR will correlate with the relative levels of exposure to that language more strongly when items are aligned with a specific language ("language-dependent") than when they are "language-independent." To make this more precise, among bilingual children, those that have more experience with English than Spanish should perform better on English non-words than their peers with less English experience. Preliminary results of an ongoing meta-analysis suggest significant associations between exposure to a given language and performance in both language-dependent and language-independent NWR (Farabolini, Taboh, Ceravolo, & Guerra, 2021). In any case, this line of research focuses on links between exposure to a given language and NWR performance. In contrast, when we discuss linguistic or cultural differences here, we ask the question of whether children vary in their performance as a function of which language they are learning (e.g., the language's typological properties) and/or their overall, absolute levels of language experience (not relative levels in a multilingual setting).

in preparation). Thus, with respect to our first research goal, Yélî Dnye is a good language to use because its large phonological inventory includes sounds that vary in cross-linguistic frequency (including some rare sounds) that can be compared in the NWR setting.

To provide some qualitative information on this inventory, we add the following observations. With only four primary places of articulation (bilabial, alveolar, post-alveolar, and velar) and no voicing contrasts, the phonological inventory is remarkably packed with acoustically similar segments. The core oral stop system includes both singleton (/p/, /t/, /ʈ/, and /k/) and doubly-articulated (/tp/, /ʈp/, /kp/) segments, with a complete range of nasal equivalents (/m/, /n/, /ɳ/, /ŋ/, /nm/, /ɳm/, /ŋm/), and with a substantial portion of them contrastively pre-nasalized or nasally released (/mp/, /nt/, /ɳʈ/, /ŋk/, /nmtp/, /ɳmʈp/, /ŋmkp/, /ʈɳ/, /kŋ/, /ʈpɳm/, /kpŋm/).[2] A large number of this combinatorial set can further be contrastively labialized, palatalized on release, or both (e.g., /p^{j}/, /p^{w}/, /p^{jw}/, /tp^{j}/, /ɳmɖb^{j}/, see Levinson, 2021 for details).

The consonantal inventory also includes a number of non-nasal continuants (/w/, /j/, /ɣ/, /l/, /β^{j}/, /l^{j}/, /lβ^{j}/). Vowels in Yélî Dnye may be oral or nasal, short or long. The 10 oral vowel qualities, which span four levels of vowel height, (/i/, /ɯ/, /u/, /e/, /o/, /ə/, /ɛ/, /ɔ/, /æ/, /ɑ/) can be produced as short and long vowels, with seven of these able to occur as short and long nasal vowels as well (/ĩ/, /ũ/, /ə̃/, /ɛ̃/, /ɔ̃/, /æ̃/, /ɑ̃/).

Our second research goal is to measure the effect of non-word length on NWR, which may need to be interpreted taking into account typical word length in the language. We estimated word length in words found in a conversational corpus (see Stimuli section for details), where the distribution of length was: 15% monosyllabic, 39% disyllabic, 29% trisyllabic, and the remaining 17% being longer than that. The vast majority of syllables use a CV format. A small portion of the lexicon features words with a final CVC syllable, but these are limited to codas of -/m/, -/p/, or -/j/ (e.g., ndap /ɳʈæp/ 'Spondylus shell') and are often resyllabified with an epenthetic /ɯ/ in spontaneous speech (e.g., ndapî /'ɳʈæpɯ/). There are also a handful of words starting with /æ/ (e.g., ala /æ'læ/ 'here') and a small collection of single-vowel grammatical morphemes (see Levinson, 2021 for details).

Our knowledge of Yélî language development is growing (e.g., Brown, 2011, 2014; Brown & Casillas, in press; Casillas, Brown, & Levinson, 2021; Liszkowski, Brown, Callaghan, Takada, & de Vos, 2012), but research into Yélî phonological development has only just begun. For example, Peute and Casillas (In preparation) find that Yélî Dnye-learning children's early spontaneous consonant productions appear to exclusively feature simplex and typologically frequent phones. Other ongoing work on Yélî Dnye includes experiment-based infant phoneme discrimination data and errors

[2] We use Levinson's (2021) under-dot notation (e.g., /ʈ/) to denote the post-alveolar place of articulation; these stops are, articulatorily, somewhat variable in place, with at least some tokens produced fully sub-apically. In approximating cross-linguistic segment frequency below we use the corresponding retroflex for each stop segment (e.g., /ʈ/, /ʈp/, /ɳ/).

made in elicited and spontaneous speech from young children, but these data are neither finalized nor yet externally reviewed (see Hellwig, Sarvasy, & Casillas, provisionally accepted for more information). These data will help better inform our current analyses based on NWR in the future (e.g., regarding common sound substitutions) but are not critical for addressing our question about the general correlation between cross-linguistic phone frequency and NWR performance.

Before closing this section, it bears mentioning that the language has an established orthography, which includes distinct graphemes for all the contrasts on which our items are based. Some children in our sample will have started school. Reading and writing instruction is currently done only in English (other than writing one's name). This was probably not the case for the majority of mothers of the children in our sample, who will have learned to read and write in Yélî Dnye during their first three years at school. It is possible that there is also some home teaching of Yélî reading and writing, notably for reading the bible.

The Yélî Community

Some aspects of the community are relevant for contextualizing our study design and results, particularly regarding sources of individual variation. Specifically, we investigated potential correlations with age, child sex, maternal education, and birth order. There is nothing particular to note regarding age and child sex, but we have some comments that pertain to the other two factors.

The typical household in our dataset includes seven individuals (typically, a mixed-sex couple and children—their own and possibly some others staying with them, as discussed in the next paragraph) and is situated among a collection of four or more other households, with structures often arranged around an open grassy area. These household clusters are organized by patrilocal relation, such that they typically comprise a set of brothers, their wives and children, and their mother and father, with neighboring hamlets also typically related through the patriline. Land attribution for building one's home is decided collectively based on land availability.

Most Yélî parents are swidden horticulturalists, who occasionally fish. Within a group of households, it is often the case that older adolescents and adults spend their day tending to their farm plots (which may not be nearby), bringing up water from the river, washing clothes, preparing food, and engaging in other such activities. Starting around age two years, children more often spend large swaths of their day playing, swimming, and foraging for fruit, nuts, and shellfish in large (~10 members) independent and mixed-age child play groups (Brown & Casillas, in press; Casillas et al., 2021). Formal education is a priority for Yélî families, and many young parents have themselves pursued additional education beyond what is locally available (Casillas et al., 2021). Local schools are well out of walking distance for many children (i.e., more than 1 hour on foot or by canoe each day), so it is very common for households situated close to a school to host their school-aged relatives during the weekdays for long segments of the school year. Children start school often at around age seven, although

the precise age depends on the child's readiness, as judged by their teacher.

Some general ideas regarding potential correlations between our NWR measures and maternal education may be drawn from the observations above. To begin with, many of our participants above 6 years of age may not be living with their birth mother but with other relatives, which may weaken associations with maternal education. In addition, it seems to us that the length of formal education a given individual may have, is not necessarily a good index of their socio-economic status or other individual properties, unlike what happens in industrialized sites. Variation may simply be due to random factors like living close to a school or having relatives there.

As for birth order, much of the work on correlations between birth order and cognitive development (including language) has been carried out in the last 70 years and in agrarian or industrialized settings (Barclay, 2015; Grätz, 2018), where nuclear families were more likely to be the prevalent rearing environment (Lancy, 2015). It is possible that birth order differences are stronger in such a setting, because much of the stimulation can only come from the parents. These effects may be much smaller in cultures where it is common for children to attend daycare at an early age (such as France) or where extended family typically live close by. The Yélî community falls in the latter case, as children are typically surrounded by siblings and cousins of several orders, regardless of their birth order in their nuclear family.

We add some observations that will help us integrate this study into the broader investigation of NWR across cultures. As mentioned previously, there is one report of relatively low NWR scores among the Tsimane', which the authors of that paper interpret as consistent with long-term effects of low levels of infant-directed speech (Cristia et al., 2020). However, Cristia et al. (2020) also point out that this is based on between-paper comparisons, and thus methods and myriad other factors have not been controlled for. The Yélî community can help us gain new insights into this matter because direct speech to children under 3 years is comparably infrequent in this community (in fact it may be infrequent in many settings, including urban ones Bunce et al., under review). Our sample also shares other societal characteristics with the Tsimane' (e.g., the community is rural and relies on farming, children grow up in wide familial networks, Casillas et al., 2021). Although infant-directed speech has been measured in different ways among the Tsimane' and the Yélî communities, our most comparable estimates at present suggest that Tsimane' young children are spoken to about 4.2 minutes per hour (Scaff, Stieglitz, Casillas, & Cristia, under review), and Yélî children about 3.6 minutes per hour (Casillas et al., 2021). Thus, if these input quantities in early childhood relate to lower NWR scores later in life, we should observe similarly low NWR scores here as in Cristia et al. (2020).

Research Questions

After some preliminary analyses to set the stage, we perform statistical analyses to inform answers to the following questions:

- Does the cross-linguistic frequency of sounds in the stimuli predict NWR scores? Are cross-linguistically rarer sounds more often substituted by commoner sounds?
- How do NWR scores change as a function of item length in number of syllables?
- Is individual variation in NWR scores correlated with child age, sex, birth order, and/or maternal education?

Throughout these analyses and in the Discussion, we also have in mind our fourth goal, namely integrating NWR results across samples varying in language and culture. We had considered boosting the interpretational value of this evidence by announcing our analysis plans prior to conducting them. However, we realized that even preregistering an analysis would be equivocal because we would not have enough power to look at all relationships of interest; in many cases possibly not enough to detect any of the known associations, given the previously discussed variability across studies. Therefore, all analyses in the present study are descriptive and should be considered exploratory.

Methods

Participants

This study was approved as part of a larger research effort by the second author. The line of research was evaluated by the Radboud University Faculty of Social Sciences Ethics Committee (Ethiek Commissie van de faculteit der Sociale Wetenschappen; ECSW) in Nijmegen, The Netherlands (original request: ECSW2017-3001-474 Manko-Rowland; amendment: ECSW-2018-041), including the use of verbal (not written) consent. As discussed in subsection "The Yélî Community," the combination of collective child guardianship practices and common hosting of school-aged children for them to attend school is that adult consent often comes from a combination of aunts, uncles, adult cousins, and grandparents standing in for the child's biological parents. Child assent is also culturally pertinent, as independence is encouraged and respected from toddlerhood (Brown & Casillas, in press). Participation was voluntary; children were invited to participate following indication of approval from an adult caregiver. Regardless of whether they completed the task, children were given a small snack as compensation. Children who showed initial interest but then decided not to participate were also given the snack.

We tested a total of 55 children from 38 families spread across four hamlet regions. We excluded test sessions from analysis for the following reasons: refused participation or failure to repeat items presented over headphones even after coaching (N=8), spoke too softly to allow offline coding (N=5), or were 13 years old or older (N=2; we tested these teenagers to put younger children at ease). The remaining 40 children (14 girls) were aged from 3 to 10 years (M = 6.40 years, SD = 1.50 years). In terms of birth order, 6 were born first, 5 second, 2 third, 7 fourth, 5 fifth, and 1 sixth, with birth order missing for 14 children. These children were tested in a hamlet far from our research

base, and we unfortunately did not ask about birth order before leaving the site. Maternal years of education averaged 8.22 years (range 6-12 years).[3] We also note that there were 34 children only exposed to Yélî Dnye at home and 6 children exposed to Yélî Dnye plus one or more other languages at home.[4]

Stimuli

Many NWR studies are based on a fixed list of 12-16 items that vary in length between 1 and 4 syllables, often additionally varying syllable complexity and/or cluster presence and complexity, and always meeting the condition that they do not mean anything in the target language (e.g., Balladares et al., 2016; Wilsenach, 2013). We kept the same variation in item length and requirement for not being meaningful in the language, but we did not vary syllable complexity or clusters because these are vanishingly rare in Yélî Dnye. We also increased the number of items an individual child would be tested on, such that a child would get up to 23 items to repeat (other work has also used up to 24-46 items: Jaber-Awida, 2018; Kalnak et al., 2014; Piazzalunga et al., 2019), with the entire test inventory of 40 final items distributed across children. We used a relatively large number of items to explore correlations with length and phonological complexity. However, aware that this large item inventory might render the task longer and more tiresome, we split items across children. Naturally, designing the task in this way may make the study of individual variation within the population more difficult because different children are exposed to different items.

A first list of candidate items was generated during a trip to the island in 2018 by selecting simple consonants (/p/, /t/, /ṭ/, /k/, /m/, /n/, /w/, /y/) and vowels (/i/, /o/, /u/, /a/, /e/) and combining them into consonant-vowel syllables, then sampling the space of resulting possible 2- to 4-syllable sequences. Candidates were automatically removed from consideration if they appeared in the most recent dictionary (Levinson, 2021). The second author presented them orally to three local research assistants, all native speakers of Yélî Dnye, who repeated each form as they would in an NWR task and additionally let the experimenter know if the item was in fact a word or phrase in Yélî Dnye. Any item reported to have a meaning or a strong association with another word form or meaning was excluded.

A second list of candidate items was generated in a second trip to the island in 2019, when data were collected by selecting complex consonants and systematically crossing them with all the vowels in the Yélî Dnye inventory to produce consonant-vowel monosyllabic forms. As before, items were automatically excluded if they appeared

[3] We asked for mothers' highest completed level of education. We then recorded the number of years entailed by having completed that level under ideal conditions.

[4] Most speakers of Yélî Dnye grow up speaking it monolingually until they begin attending school around the age of 7 years; school instruction is in English. While monolingual Yélî Dnye upbringing is common, multilingual families are not unusual, particularly in the region around the Catholic Mission (the same region in which much of the current data were collected), where there is a higher incidence of married-in mothers from other islands (Brown & Casillas, in press). Children in these multilingual families grow up speaking Yélî Dnye plus English, Tok Pisin, and/or other language(s) from the region.

in the dictionary. Furthermore, since perceiving vowel length in isolated monosyllables is challenging, any item that had a short/long lexical neighbor was excluded. We made sure that the precise consonant-vowel sequence occurred in some real word in the dictionary (i.e., there existed a longer word that included the monosyllable as a sub-sequence). These candidates were then presented to one informant, for a final check that they did not mean anything. Together with the 2018 selection, they were recorded, based on their orthographic forms, using a Shure SM10A XLR dynamic headband microphone and an Olympus WS-832 stereo audio recorder (using an XLR to mini-jack adapter) by the same informant, and monitored by the second author for clear production of the phonological target. The complete recorded list was finally presented to two more informants, who were able to repeat all the items and who confirmed there were no real words present. Despite these checks, one monosyllable was ultimately frequently identified as a real word in the resulting data (intended yî /yɯ/; identified as yi /yi/, 'tree'). Additionally, an error was made when preparing files for annotation, resulting in two items being merged (tpâ /tpɑ/ and tp:a /tpæ̃/). These three problematic items are not described here, and are removed from the analyses below.

The final list includes three practice items and 40 test items (across children): 16 monosyllables containing sounds that are less frequent in the world's languages than singleton plosives; 8 bisyllables; 12 trisyllables; and 4 quadrisyllables (see Table 1).

Table 1. ***NWR stimuli in orthographic (Orth.) and phonological (Phon.) representations, as a function of item type.***

Practice		**Monosyllabic**		**Bisyllabic**		**Trisyllabic**		**Tetrasyllabic**	
Orth.	**Phon.**	**Orth.**	**Phon.**	**Orth.**	**Phon.**	**Orth.**	**Phon.**	**Orth.**	**Phon.**
nopimade	nɔpimæt̪ɛ	dp:a	t̪pæ̃	kamo	kæmɔ	dimope	t̪imɔpɛ	dipońate	t̪ipɔnætɛ
poni	pɔni	dpa	t̪pæ	kańi	kæni	diyeto	t̪ijɛtɔ	ńomiwake	nɔmiwækɛ
wî	wɯ	dpâ	t̪pɑ	kipo	kipɔ	meyadi	mɛjæt̪i	todiwuma	tɔt̪iwumæ
		dpê	t̪pə	ńoki	nɔki	mituye	mitujɛ	wadikeńo	wæt̪ikɛnɔ
		dpéé	t̪peː	ńomi	nɔmi	ńademo	næt̪ɛmɔ		
		dpi	t̪pi	piwa	piwæ	ńayeki	næjɛki		
		dpu	t̪pu	towi	tɔwi	ńuyedi	nujɛt̪i		
		gh:ââ	ɣɑ̃ː	tupa	tupæ	pedumi	pɛt̪umi		
		ghuu	ɣuː			tiwuńe	tiwunɛ		
		kp:ââ	kpɑ̃ː			tumowe	tumɔwɛ		

kpu	kpu	widoǹe	wiṭɔnɛ
lv:ê	lβʲə̃	wumipo	wumipɔ
lva	lβʲæ		
lvi	lβʲi		
t:êê	tə̃ː		
tpê	tpə		

A Praat script (Boersma & Weenink, 2020) was written to randomize this list 20 times, and to split it into two sub-lists, to generate 40 different elicitation sets. The 40 elicitation sets are available online from osf.io/dtxue/. The split had the following constraints:

- The same three items were selected as practice items and used in all 40 elicitation sets.
- Splits were done within each length group from the 2018 items (i.e., separately for 2-, 3-, and 4-syllable items); and among onset groups for the difficult monosyllables generated in 2019 (i.e., all the monosyllables starting with /tp/ were split into 2 sub-lists). Since some of these groups had an odd number of items, one of the sub-lists was slightly longer than the other (20 vs. 23).
- Once the sub-list split had been done, items were randomized such that all children heard first the 3 practice items in a fixed order (1, 2, and 4 syllables), a randomized version of their sub-list selection of difficult onset items, and randomized versions of their 2-syllable, then 3-syllable, and finally 4-syllable items.

Cross-linguistic Frequency

To inform our analyses, we estimated the typological frequency of all phonological segments present in the target items using the PHOIBLE cross-linguistic phonological inventory database (Moran & McCloy, 2019). For each phone in our task, we extracted the number and percentage of languages noted to have that phone in its inventory. While PHOIBLE is unprecedented in its scope, with phonological inventory data for over 2000 languages at the time of writing, it is of course still far from complete, which may mean that frequencies are estimates rather than precise descriptors. Note that nearly half of the phones in PHOIBLE are only attested in one language (Steven Moran, personal communication). Extrapolating from this observation, we treat the three segments in our stimuli that were unattested in PHOIBLE (/lβʲ/, /ṭp/, and /tp/) as having a frequency of 1 (i.e., appearing in one language), with a (rounded) percentile of 0% (i.e., its cross-linguistic percentile is zero).

Within-language Frequency

Additionally, we estimated the usage frequency of the phones present in the target items in a corpus of child-centered recordings (Casillas et al., 2021). That corpus was

constituted by sampling from audio-recordings (7–9 hours long), collected as 10 children aged between 1 month and 3 years went about their day. The researchers selected 9 2.5-minute clips randomly and 11 1- or 5-minute clips by hand (selected to represent peak turn-taking and child vocal activity). These clips were segmented and transcribed by the lead researcher and a highly knowledgeable local assistant, who speaks Yélî Dnye natively, has ample experience in this kind of research, and often knew all the recorded people personally. For more details, please refer to Casillas et al. (2021).

For the present study, we extracted the transcriptions of adult speech (i.e., removing key child and other children's speech) and split them into words using white space. We then removed all English and Tok Pisin words. The resulting corpus contained a total of 18,934 word tokens of 1,686 unique word types. To get our phone frequency measure, we counted the number of word types in which the phone occurred, and applied the natural logarithm.[5] Here, unattested sounds were not considered (i.e., they were declared NA so that they do not count for analyses). Note that the resulting values estimate usage frequencies for very young children's input and, while this is somewhat different from what our older participants experience on a daily basis, we can expect that this is a reasonable approximation of the early input that formed the foundation of their phonological knowledge.

Procedure

There is some variation in procedure in previous work. For example, while items are often presented orally by the experimenter (Torrington Eaton, Newman, Ratner, & Rowe, 2015), an increasing number of studies have turned instead to playing back prerecorded stimuli in order to increase control in stimulus presentation (Brandeker & Thordardottir, 2015).

In adapting the typical NWR procedure for our context, we balanced three desiderata: That children would not be unduly exposed to the items before they themselves had to repeat them (i.e., from other children who had participated); that children would feel comfortable doing this task with us; and that community members would feel comfortable having their children do this task with us.

We tested in four different sites spread across the northeastern region of the island, making a single visit to each, conducting back-to-back testing of all eligible children present at the time of our visit in order to prevent the items from 'spreading' between children through hearsay. Whenever children living in the same household were tested, we tried to test children in age order, from oldest to youngest, to minimize intimidation for younger household members, and always using different elicitation sets. Because space availability was limited in different ways from hamlet to hamlet, the places where elicitation happened varied across testing sites. More information is

[5] We also carried out analyses using token (rather than type) phone frequency, but this measure was not correlated with whole-item NWR scores, and therefore the fact that it did not explain away the predictive value of cross-linguistic phone frequency was less informative than the relationship discussed in the Results section.

available from the online materials (https://osf.io/qt8gr/).

We tested one child at a time. We fitted the child with a headset microphone (Shure SM10A or WH20 XLR with a dynamic microphone on a headband, most children using the former) that fed into the left channel of a Tascam DR40x digital audio recorder. The headsets were designed for adult use and could not be comfortably seated on many children's heads without a more involved adjustment period. To minimize adjustment time, which was uncomfortable for some children given the proximity of the foreign experimenter and equipment, we placed the headband on children's shoulders in these cases, carefully adjusting the microphone's placement so that it was still close to the child's mouth. A research assistant who spoke Yélî Dnye natively, and who could also hear the instructions over headphones, sat next to the child throughout the task to provide instructions and, if needed, encouragement. The research assistant coached the child throughout the task to make sure that they understood what they were expected to do. Finally, an experimenter (the first author) was also fitted with headphones and a microphone. She was in charge of delivering the pre-recorded stimuli to the research assistant, the child, and herself over headphones.

The first phase of the experiment involved making sure the child understood the task. We explained the task and then presented the first practice item. At this point, many children did not say anything in response, which triggered the following procedure: First, the assistant insisted the child make a response. If the child still did not say anything, the assistant said a real word and then asked the child to repeat it, then another and another. If the child could repeat real words correctly, we provided the first training item over headphones again for children to repeat. Most children successfully started repeating the items at this point, but a few needed further help. In this case, the assistant modeled the behavior (i.e., the child and assistant would hear the item again, and the assistant would repeat it; then we would play the item again and ask the child to repeat it). A small minority of children still failed to repeat the item at this point. If so, we tried again with the second training item, at which point some children demonstrated task understanding and could continue. A fraction of the remaining children, however, failed to repeat this second training item, as well as the third one, in which case we stopped testing altogether (see Participants section for exclusions).

The second phase of the experiment involved going over the list of test items randomly assigned to each child. This was done in the same manner as the practice items: the stimulus was played over the headphones, and then the child repeated it aloud. NWR studies vary in whether children are allowed to hear and/or repeat the item more than one time. We had a fixed procedure for the test items (i.e., the non-practice items) in which the child was allowed to make further attempts if their first attempt was judged erroneous in some way by the assistant. The procedure worked as follows: When the child made an attempt, the assistant indicated to the experimenter whether the child's production was correct or not. If correct, the experimenter would whisper this note of correct repetition into a separate headset that fed into the right channel

of the same Tascam recorder and we moved on to the next item. If not, the child was allowed to try again, with up to five attempts allowed before moving on to the next item. Children were not asked to make repetitions if they did not produce a first attempt. In total, the sessions took approximately six minutes (one for practice; five for the test list).

Coding

The first author then annotated the onset and offset of all children's productions from the audio recording using Praat audio annotation software (Boersma & Weenink, 2020), then ran a script to extract these tokens, pairing them with their original auditory target stimulus, and writing these audio pairs out to .wav clips. The assistant then listened through all these paired target-repetition clips randomized across children and repetitions, grouped such that all the clips of the same target were listened to in succession. For each clip, the assistant indicated in a notebook whether the child production was a correct or incorrect repetition and orthographically transcribed the production, noting when the child uttered a recognizable word or phrase and adding the translation equivalent of that word/phrase into English. The assistant was also provided with some general examples of the types of errors children made without making specific reference to Yélî sounds or the items in the elicitation sets. Because the phonological inventory is so acoustically packed and annotation was done based on audio data alone, it might be easy to misidentify a segment. Therefore, the assistant double-checked all of her annotations by listening to them and assessing them a second time, once she had completed a full first round.

Analyses

Previous work typically reports two scores: a binary word-level exact repetition score, and a phoneme-level score, defined as the number of phonemes that can be aligned across the target and attempt, divided by the number of phonemes of whichever item was longer (the target or the attempt; as in Cristia et al., 2020). Previous work does not use distance metrics, but we report these rather than the phoneme-level scores because they are more informative. To illustrate these scores, recall our example of an English target being /bilik/ with an imagined response [bilig]. We would score this response as follows: at the whole item level this production would receive a score of zero (because the repetition is not exact); at the phoneme level this production would receive a score of 80% (4 out of 5 phonemes repeated exactly); and the phone-based Levenshtein distance for this production is 20% (because 20% of phonemes were substituted or deleted). Notice that the phone-based Levenshtein distance is the complement of the phoneme-level NWR score. An advantage of using phone-based Levenshtein distance is that it is scored automatically with a script, and it can then easily be split in terms of deletions and substitutions (insertions were not attested in this study).

Results

Preliminary Analyses

We first checked whether whole-item NWR scores varied between first and subse-

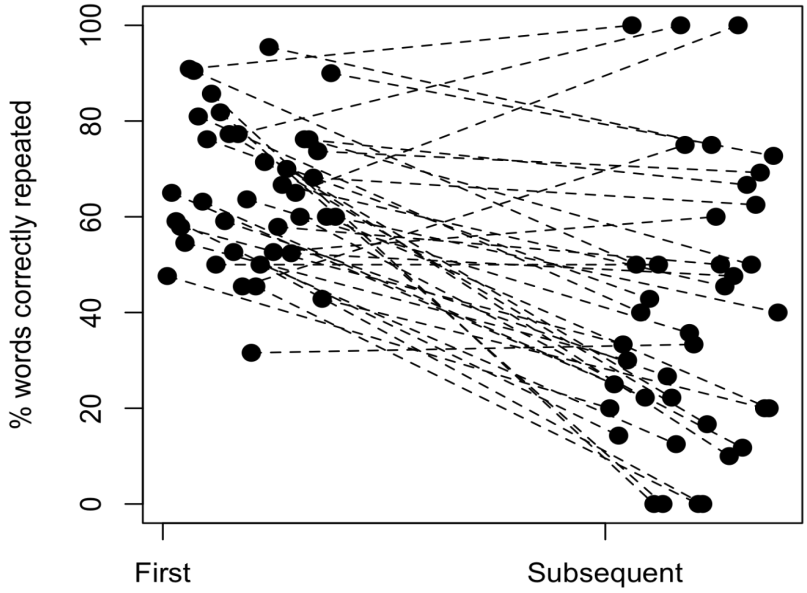


Figure 1. ***Whole-item NWR scores for individual participants averaging separately their first attempts and all other attempts.***

quent presentations of an item by averaging word-level scores at the participant level separately for first attempts and subsequent repetitions. We excluded 1 child who did not have data for one of these two types. As shown in Figure 1, participants' mean word-level scores became more heterogeneous in subsequent repetitions. Surprisingly, whole-item NWR scores for subsequent repetitions (M = 40, SD = 28) were on average lower than first ones (M = 65, SD = 15), $t(38) = 5.89$, $p < 0.001$; Cohen's d = 1.13). Given uncertainty in whether previous work used first or all repetitions, and given that scores here declined and became more heterogeneous in subsequent repetitions, we focus the remainder of our analyses only on first repetitions, with the exception of qualitative analyses of substitutions.

Taking into account only the first attempts, we derived overall averages across all items. The overall NWR score was M = 65% (SD = 15%), Cohen's d = 4.39. The phoneme-based normalized Levenshtein distance was M = 21% (SD = 9%), meaning that about a fifth of phonemes were substituted or deleted.

We also looked into the frequency with which mispronunciations resulted in real words. In fact, two thirds of incorrect repetitions were recognizable as real words or phrases in Yélî Dnye or English: 63%. This type of analysis is seldom reported. We could only find one comparison point: Castro-Caldas, Petersson, Reis, Stone-Elander, and Ingvar (1998) found that illiterate European Portuguese adults' NWR mispronunciations resulted in real words in 11.16% of cases, whereas literate participants did so

in only 1.71% of cases. The percentage we observe here is much higher than reported in the study by Castro and colleagues, but we do not know whether age, language, test structure, or some other factor explains this difference, such as the particularities of the Yélî Dnye phonological inventory, which lead any error to result in many true-word phonetic neighbors. Follow-up work exploring this type of error in children from other populations in addition to further work on Yélî children may clarify this association.

NWR and Typology: NWR as a Function of Cross-Linguistic Phone Frequency

Turning to our first research question, we analyzed variation in whole-item NWR scores as a function of the average frequency with which sounds composing individual target words are found in languages over the world. To look at this, we fit a mixed logistic regression in which the outcome variable was whether the non-word was correctly repeated or not. The fixed effect of interest was the average cross-linguistic phone frequency; we also included child age as a control fixed effect, in interaction with cross-linguistic phone frequency, and allowed intercepts to vary over the random effects child ID and target ID.

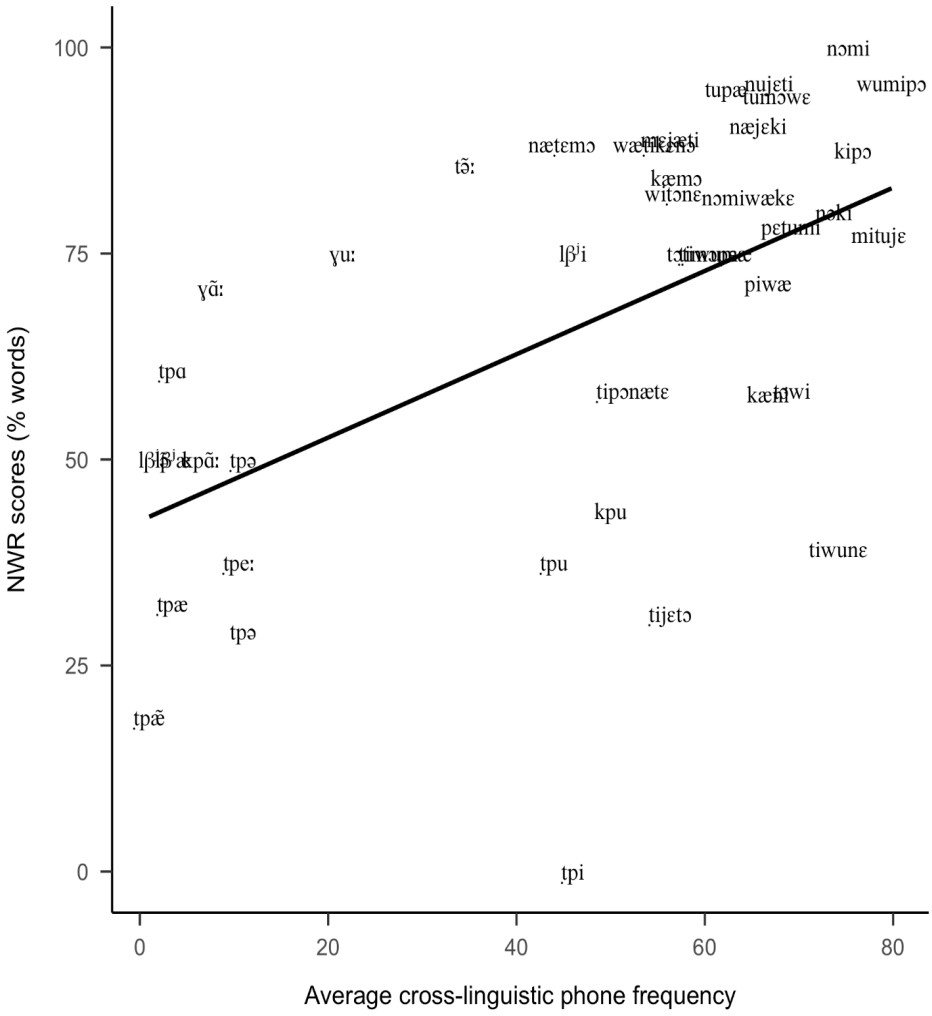


Figure 2. ***NWR scores for individual target words as a function of the average frequency with which each phone is found across languages.***

We could include 826 observations, from 40 children producing in any given trial one of 40 potential target words. The analysis revealed a main effect of age (ß = 0.39, SE ß = 0.13, $p < 0.01$), with older children repeating more items correctly. It also revealed a significant estimate for the scaled average cross-linguistic frequency of phones in the target words (ß = 0.80, SE ß = 0.19, $p < 0.001$): Target words with phones found more frequently across languages had higher correct repetition scores, as shown in Figure 2. Averaging across participants, the Pearson correlation between scaled average cross-linguistic phone frequency and whole-item NWR scores was $r(38) = .544$.

Additionally, the effect for the interaction between the two fixed effects was small but significant (ß = 0.22, SE ß = 0.09, $p = 0.01$): The effect of frequency was larger for older children. Inspection of Figure 3 suggests that the age effects are more marked for items containing cross-linguistically common phones, such that children's average performance increases more rapidly with age for those than for items containing cross-linguistically uncommon phones.

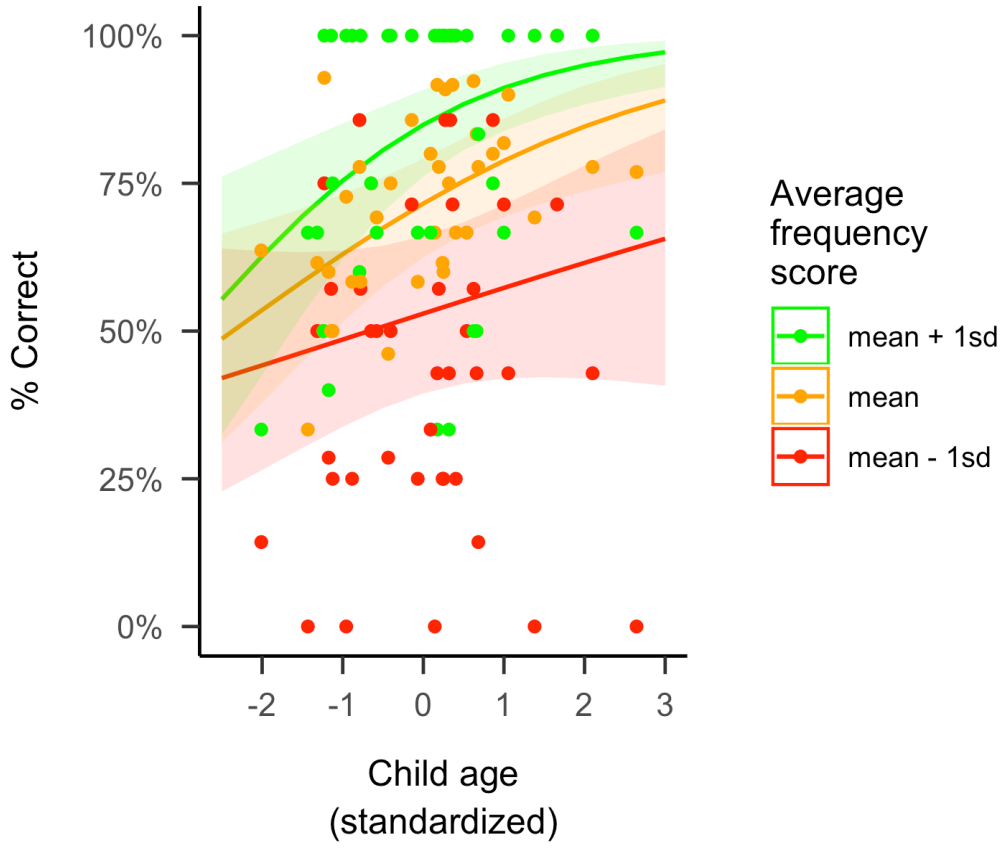


Figure 3. ***NWR scores as a function of age and typological frequency. Lines are fits from the model in the main text predicting NWR scores from child age (x axis) and the average frequency with which each phone is found across languages (mean, or plus/minus one standard deviation). Each circle indicates the estimated NWR scores***

for one child at one frequency level.

NWR and Typology: NWR as a Function of Within-Language Phone Frequency

We next checked whether the association between whole-item NWR scores and cross-linguistic phone frequency could actually be due to frequency of the sounds within the language: The same perception and production pressures that shape languages diachronically could affect a language's lexicon, so that sounds that are easier to perceive or produce are more frequent within a language than those that are harder. If so, children will have more experience with the easier sounds, and they may thus be better able to represent and repeat non-words containing them simply because of the additional exposure.

Phone corpus-based frequencies were correlated with phone cross-linguistic frequencies [$r(27)=0.50$, $p < 0.01$]; and item-level average phone corpus-based frequencies were correlated with the corresponding cross-linguistic frequencies [$r(38)=0.73$, $p < 0.001$]. Moreover, averaging across participants, the Pearson correlation between scaled average corpus phone frequency and whole-item NWR scores was $r(38)=.432$, $p < 0.01$. Therefore, we fit another mixed logistic regression, this time declaring as fixed effects both scaled cross-linguistic and corpus frequencies (averaged across all attested phones within each stimulus item), in addition to age. As before, the model contained random slopes for both child ID and target. In this model, both cross-linguistic phone frequency (ß = 0.78, SE ß = 0.27, $p < 0.01$) and age (ß = 0.35, SE ß = 0.13, $p < 0.01$) were significant predictors of whole-item NWR scores, but corpus phone frequency (ß = 0.00, SE ß = 0.25, $p = 0.99$) was not.

Follow-up Analyses: Patterns in NWR Mispronunciations.

We addressed our first research question in a second way, by investigating patterns of error. Unlike all other analyses, we looked at all attempts, so as to base our generalizations on more data. As in all analyses, we did not exclude errors resulting in real words. Deletions were very rare (insertion and metathesis were not attested): there were only 17 instances of deleted vowels (~0.35% of all vowel targets), and 13 instances of deleted consonants (~0.50% of all consonant targets). We therefore focus our qualitative description here on substitutions: There were 813 cases of substitutions, ~16.81 of the 4836 phones found collapsing across all children and target words, so that substitutions constituted the majority of incorrect phones (~96.10% of unmatched phones). To inform our understanding of how cross-linguistic patterns may be reflected in NWR scores, we asked: Is it the case that cross-linguistically less common and/or more complex phones are more frequently mispronounced, and more frequently substituted by more common ones than vice versa?[6]

We looked for potential asymmetries in errors for different types of sounds in vowels

[6] Note that tables of errors including child age are provided in the project repository for those interested in a finer-grained analysis than what is presented here. See https://osf.io/5qspb/wiki/home/, quick links, error tables.

by looking at the proportion of vowel phones that were correctly repeated or not, generating separate estimates for nasal and oral vowels. The nasal vowels in our stimuli occur in ~1.40% of languages' phonologies (range 0% to 3%); whereas oral vowels in our stimuli occur in ~31.55% of languages' phonologies (range 3% to 92%). As noted above, frequency within the language is correlated with cross-linguistic frequency, and thus these two types of sounds also differ in the former: Their frequencies in Yélî Dnye are: nasal vowels ~0.03‰ (range 0.00‰ to 0.05‰) versus oral ~0.23‰ (range 0.02‰ to 0.76‰).

We distinguished errors that included a change of nasality (and may or may not have preserved quality), versus those that preserved nasality (and were therefore a quality error), shown in Table 2. We found that errors involving nasal vowel targets were more common than those involving oral vowels (35.70 versus 12.10%). Additionally, errors in which a nasal vowel lost its nasal character were 10 times more common than those in which an oral vowel was produced as a nasal one. Note that this analysis does not tell us whether cross-linguistic or within-language frequency is the best predictor, an issue to which we return below.

Table 2. ***Number (and percent) of vowel targets that were correctly repeated (Corr.), deleted (Del.), or substituted, as a function of vowel type, and whether the error resulted in a nasality change (Nasal Err.) or only a quality change (Qual. Err.)***

	Corr.	**Del.**	**Nasal Err.**	**Qual. Err.**	**% Corr.**	**% Del.**	**% Nasal Err.**	**% Qual Err.**
Nasal Target	101	0	39	17	64.3	0	24.8	10.8
Oral Target	1988	17	52	204	87.9	0.8	2.3	9

For consonants, we inspected complex ([t̪p], [tp], [kp], [km], [kn̪], [mp], and [lβʲ]) versus simpler ones ([m], [n], [l], [w], [j], [w], [t̪], [g], [p], [t], [k], [f], [ɣ], [h], and [tʃ]), using the same logic: We looked at correct phone repetition, substitution with a change in complexity category, or a change within the same complexity category.[7] The complex consonants in our stimuli occur in ~17.33% of languages 'phonologies (range 0% to 78%); whereas simple consonants in our stimuli occur in ~67.62% of languages 'phonologies (range 13% to 96%). Again these groups of sounds differ in their frequency within the language. Their type frequencies in Yélî Dnye are: complex consonants ~0.04‰ (range 0.00‰ to 0.10‰) versus simple consonants ~0.32‰ (range 0.06‰ to 0.55‰).

Table 3 showed that errors involving complex consonant targets were more common than those involving simple consonants (57 versus 8.20%). Additionally, errors in which a complex consonant was mispronounced as a simple consonant were quite common, whereas those in which a simple consonant was produced as a complex one

[7] Note that the substitutions included phones that are not native to Yélî Dnye but do occur in English (e.g., [tʃ]). These data come from careful transcriptions by a native Yélî Dnye speaker who is very fluent in English.

were vanishingly rare.

Table 3. ***Number (and percent) of consonant targets that were correctly repeated (Corr.), deleted (Del.), or substituted, as a function of the complexity of the consonant, and whether the error resulted in a change of complexity (Cmpl Err.) or not (Othr Err.)***

	Corr.	Del.	Cmpl Err.	Othr Err.	% Corr.	% Del	% Cmpl Err.	% Othr Err.
Complex Target	198	0	219	44	43	0	47.5	9.5
Simple Target	1482	13	3	117	91.8	0.8	0.2	7.2

To address whether errors were better predicted by cross-linguistic or within-language frequency, we calculated a proportion of productions that were correct for each phone (regardless of the type of error or the substitution pattern). Graphical investigation suggested that in both cases the relationship was monotonic and not linear, so we computed Spearman's rank correlations between the correct repetition score, on the one hand, and the two possible predictors on the other. Although we cannot directly test the interaction due to collinearity, the correlation with cross-linguistic frequency [$r(346.78)=0.74$, $p < 0.001$] was greater than that with within-language frequency [$r(817.23)=0.39$, $p = 0.09$].

Length Effects on NWR

We next turned to our second research question by inspecting whether NWR scores varied as a function of word length (Table 4). In this section and all subsequent ones, we only look at first attempts, for the reasons discussed previously. Additionally, we noticed that participants scored much lower on monosyllables than on non-words of other lengths. This is likely due to the fact that the majority of monosyllables were designed to include sounds that are rare in the world's languages, which may be harder to produce or perceive, as suggested by our previous analyses of NWR scores as a function of cross-linguistic phone frequency and error patterns. Therefore, we set monosyllables aside for this analysis.

Table 4. ***NWR means (and standard deviations) measured in whole-word scores and normalized Levenshtein Distance (NLD), separately for the four stimuli lengths.***

	Word	**NLD**
1 syll	48 (22)	40 (18)
2 syll	79 (22)	8 (9)
3 syll	78 (19)	7 (7)
4 syll	74 (32)	9 (12)

We observed the typical pattern of lower scores for longer items only for the whole-item scoring, and even there differences were rather small. In a generalized binomial mixed model excluding monosyllables, we included 479 observations, from 40 children producing, in any given trial, one of 24 (non-monosyllabic) potential target words. The analysis revealed a positive effect of age (ß = 0.56, SE ß = 0.14, $p < 0.001$) and a negative but non-significant estimate for target length in number of syllables (ß = -0.15, SE ß = 0.33, p = 0.65).

Individual Variation and NWR

Our final exploratory analysis assessed whether variation in scores was structured by factors that vary across individuals, as per our third research question. As shown in Figure 4, there was a greater deal of variance across the tested age range, with significantly higher NWR scores for older children (Spearman's rank correlation, given inequality of variance): $\rho(38) = .47$, $p < 0.01$. In contrast, there was no clear association between NWR scores and sex: Welch t (27.33) = -0.60, p = 0.56; NWR scores and birth order (data missing for 14 children): $\rho(24) = -.198$, p = 0.33; or NWR scores and maternal education: $\rho(38) = .097$, p = 0.55.

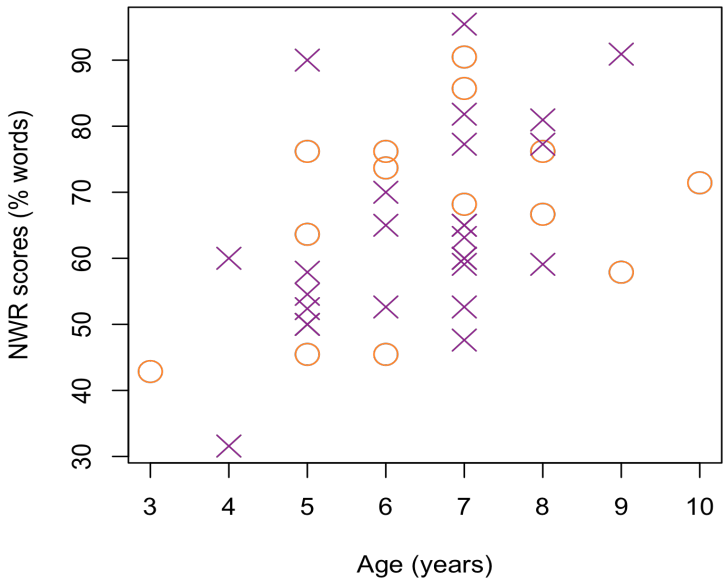


Figure 4. ***NWR whole-item scores for individual participants as a function of age and sex (purple crosses = boys, orange circles = girls).***

Discussion

We used non-word repetition to investigate phonological development in a language with a large phonological inventory (including some typologically rare segments). We aimed to provide additional data on two questions already visited in NWR work, namely the influence of stimulus length and individual variation, plus one research

area that has received less attention, regarding the possible correlation between typological phone frequency and NWR scores. An additional overarching goal was to discuss NWR in the context of population and language diversity, since it is very commonly used to document phonological development in children raised in urban settings with wide-spread literacy, and has been seldom used in non-European languages (but note there are exceptions, including work cited in the Introduction and in the Discussion below). We consider implications of our results on each of these four research areas in turn.

NWR and Typology

Arguably the most innovative aspect of our data relates to the inclusion of phones that are less commonly found across languages, and rarely used in NWR tasks. As explained in the Introduction, typological frequency of phones could reflect ease of perception, ease of production, and other factors, and these factors could affect speech processing and production. This predicts a correlation between typological frequency and NWR performance, due to those factors affecting both. To assess this prediction, we looked at our data in two ways. First, we measured the degree of association between NWR scores and cross-linguistic frequency at the level of non-word items. Second, we described mispronunciation patterns, by looking at correct and incorrect repetitions of simpler and more complex sounds, which are also more or less frequent.

There are some reasons to believe that Yélî Dnye put that hypothesis to a critical test: The phoneme inventory is both large and acoustically packed, in addition to containing several typologically infrequent (or unique) contrasts. One could then predict that correlations with typological frequency should be relatively weak because the ambient language puts more pressure on Yélî children to distinguish (perceptually and articulatorily) fine-grained phonetic differences than what is required of child speakers of other languages. On the other hand, it is also possible that this pressure gives Yélî children no benefit, and that some of these categories are simply acquired later in development. We can draw a parallel with children learning another Papuan language, Ku Waru, which has a packed inventory of lateral consonants; where children do not produce adult-like realizations of the more complex of these laterals (the prestopped velar lateral /gʟ/) until 5 or 6 years of age (Rumsey, 2017).

We do not have the necessary data to assess whether the correlation is indeed weaker for Yélî Dnye learners than learners of other languages, but we did find a robust correlation of average segmental cross-linguistic frequency and NWR performance: Even accounting for age and random effects of item and participant, we saw that target words with typologically more common segments were repeated correctly more often. This effect was large, with a magnitude more than twice the size of the effect of participant age. Additionally, we observed an interaction between age and this factor, which emerged because cross-linguistic frequency explained more variance at older ages (i.e., the difference in performance for more versus less typologically frequent

sounds was greater for older than younger children). Importantly, the correlation between performance and typological frequency remained significant after accounting for the frequencies of these segments in a conversational corpus. An analysis of the substitutions made by children also aligned with this interpretation, with typologically more common sounds being substituted for typologically less common ones.

We thus at present conclude that typological frequency of sounds is, to a certain extent, mirrored in children's NWR, in ways that may not be due merely to how often those sounds are used in the ambient language, and which are not erased by language-specific pressure to make finer-grained differences early in development. We do not aim to reopen a debate on the extent to which cross-linguistic frequency of occurrence can be viewed necessarily as reflecting ease of perception or production (via phonotactic constraints, ambiguous parsing conditions, individual differences, and more as in, e.g., Beddor, 2009; Bermúdez-Otero, 2015; Maddieson, 2009; Ohala, 1981; Yu, 2021), but we do point out that this association is interestingly different from effects found in artificial language learning tasks (see Moreton & Pater, 2012 for a review) which are in some ways quite similar to NWR. We believe that it may be insightful to extend the purview of NWR from a narrow focus on working memory and structural factors to broader uses, including for describing the phonological representations in the perception-production loop (as in e.g., Edwards, Beckman, & Munson, 2004).

Length Effects and NWR

We investigated the effect of item complexity on NWR scores by varying the number of syllables in the item. In broad terms, children should have higher NWR scores for shorter items. That said, previous work summarized in the Introduction has shown both very small (e.g., Piazzalunga et al., 2019) and very large (e.g., Cristia et al., 2020) effects of stimulus length. Setting aside our monosyllabic stimuli (which contained typologically infrequent segments with lower NWR scores, as just discussed), we examined effects of item length among the remaining stimuli, which range between 2 and 4 syllables long. The effect of item length was not significant in a statistical model that additionally accounted for age and random effects of item and participant. We do not have a good explanation for why samples in the literature vary so much in terms of the size of length effects, but two possibilities are that this is not truly a length effect but a confound with some other aspect of the stimuli, or that there is variation in phonological representations that is poorly understood. We explain each idea in turn.

First, it remains possible that apparent length effects are actually due to uncontrolled aspects of the stimuli. For instance, some NWR researchers model their non-words on existing words, by changing some vowels and consonants, which could lead to fewer errors (since children have produced similar words in the past); some researchers control tightly the diphone frequency of sub-sequences in the non-words. Building on these two aspects that researchers often control, one can imagine that longer items have fewer neighbors, and thus both the frequency with which children have

produced similar items and (relatedly) their n-phone frequency is overall lower. If this idea is correct, a careful analysis of non-words used in previous work may reveal that studies with larger length effects just happened to have longer non-words with lower n-phone frequencies.

Second, NWR is often described as a task that tests flexible perception-production, and as such it is unclear why length effects should be observed at all. However, it is possible that NWR relies on more specific aspects of perception-production, in ways that are dependent on stimulus length. A hint in this direction comes from work on illiterate adults, who can be extremely accurate when repeating short non-words, but whose NWR scores are markedly lower for longer items. In a longitudinal study on Portuguese-speaking adults who were learning to read, Kolinsky, Leite, Carvalho, Franco, and Morais (2018) found that, before reading training, the group scored 12.5% on 5-syllable items, whereas after 3 months of training, they scored 62.5% on such long items, whereas performance was at 100% for monosyllables throughout. Given that as adults they had fully acquired their native language, and obviously they had flexible perception-production schemes that allowed them to repeat new monosyllables perfectly, the change that occurred in those three months must relate to something else in their phonological skills, something that is not essential to speak a language natively. Thus, we hazard the hypothesis that sample differences in length effects may relate to such non-essential skills. Since as stated this hypothesis is underspecified, further conceptual and empirical work is needed.

Individual Variation and NWR

Our review of previous work in the Introduction suggested that our anticipated sample size would not be sufficient to detect most individual differences using NWR. We give a brief overview of individual difference patterns of four types in the present data—age, sex, birth order, and maternal education—hoping that these findings can contribute to future meta- or mega-analytic efforts aggregating over studies.

In broad terms, we expected that NWR scores would increase with participant age, as this is the pattern observed in several previous studies (English Vance et al., 2005; Italian Piazzalunga et al., 2019; Cantonese Stokes et al., 2006; but not in Cristia et al., 2020). Indeed, age was significantly correlated with NWR scores and it also showed up as a significant predictor of NWR score when included as a control factor in the analyses of both item length and average segmental frequency. In brief, our results underscore the idea that phonological development continues well past the first few years of life, extending into middle childhood and perhaps later (Hazan & Barrett, 2000; Rumsey, 2017).

In contrast, previous work varies with respect to correlations of NWR scores with maternal education (e.g., Farmani et al., 2018; Kalnak et al., 2014; Meir & Armon-Lotem, 2017). We did not expect large correlations with maternal education in our sample for two reasons: First, education on Rossel Island is generally highly valued and so wide-

spread that little variation is seen there; second, formal education is not at all essential to ensuring one's success in society and may not be a reliable index of local socioeconomic variation. In fact, maternal education correlated with NWR score at about r~.1, which is small. We find correlations of about that size for participant sex, which is aligned with previous work (Chiat & Roy, 2007).

Finally, we investigated whether birth order might correlate with NWR scores, as it does with other language tasks, such that first-born children showing higher scores on standardized language tests than later-born children (Havron et al., 2019) and adults (in a battery including verbal abilities, e.g., Barclay, 2015), presumably because later-born children receive a smaller share of parental input and attention than first-borns. Given shared caregiving practices and the hamlet organization typical of Rossel communities, children have many sources of adult and older child input that they encounter on a daily basis and first-born children quickly integrate with a much larger pool of both older and younger children with whom they partly share caregivers. Therefore we expected that any correlations with birth order on NWR would be attenuated in this context. In line with this prediction, our descriptive analysis showed a non-significant correlation between birth order and NWR score. However, the effect size was larger than that found for the other two factors and it is far from negligible, at r~.2 or Cohen's d~0.41. In fact, two large studies (with therefore precise estimates) found effects of about d~.2 for birth order effects on other language tasks (Barclay, 2015; Havron et al., 2019), which would suggest the correlations we found are larger. We therefore believe it may be worth revisiting this question with larger samples in similar child-rearing environments, to further assess whether distributed child care results in more even language outcomes for first- and later-born children.

NWR across Languages and Cultures

The fourth research area to which we wanted to contribute pertained to the use of NWR across languages and populations, since when designing this study we wondered whether NWR was a culture-fair test of phonological development. Although our data cannot answer this question because we have only sampled one language and population here, we would like to spend some time discussing the integration of these results to the wider NWR literature. It is important to note at the outset that we cannot obtain a final answer because integration across studies implies not only variation in languages and child-rearing settings, but also in methodological aspects including non-word length, non-word design (e.g., the syllable and phone complexity included in the items), and task administration, among others. Nonetheless, we feel the NWR task is prevalent enough to warrant discussion about this, similarly to other tasks sometimes used to describe and compare children's language skills across populations, like the recent re-use of the MacArthur-Bates Communicative Development Inventory to look at vocabulary acquisition across multiple languages (Frank et al., 2017).

The range of performance we observed overlapped with previously observed levels of performance. Paired with our thorough training protocol, we had interpreted the

NWR scores among Yélî Dnye learners as indicating that our adaptations of NWR for this context were successful, even given a number of non-standard changes to the training phase and to the design of the stimuli. Additionally, it seemed that Yélî children showed comparable performance to others tested on a similar task, despite the many linguistic, cultural, and socioeconomic differences between this and previously tested populations, unlike the case that had been reported for the Tsimane' (Cristia et al., 2020).

Comparison across published studies is difficult (see SM2 for our preliminary attempt). To be certain whether language-specific characteristics do account for meaningful variation in NWR scores, it will be necessary to design NWR tasks that are cross-linguistically valid. We believe this will be exceedingly difficult (or perhaps impossible), since it would entail defining a 10-20 set of items that are meaningless, but phonotactically legal, in all of the languages. An alternative may be to find ways to regress out some of these differences, and thus compare languages while controlling for choices of phonemes, syllable structure, and overall length of the NWR items. Both of these issues are discussed in Chiat (2015). As for the variable strengths of age correlations discussed above, here as well we are uncertain to what they may be due, but we do hope that these intriguing observations will lead others to collect and share NWR data.

Limitations

Before closing, we would like to point out some salient limitations of the current work. To begin with, we only employed one set of non-words, in which not all characteristics that previous work suggest matter were manipulated (Chiat, 2015). As a result, we only have a rather whole-sale measure of performance, and we do not know to what extent lexical knowledge, pure phonological knowledge, and working memory, among others, contribute to children's performance. Similarly, our items varied systematically in length and typological frequency of the sounds included, but not in other potential dimensions (such as whether the items contained morphemes of the language or not).

We relied on a single resource, PHOIBLE, for our estimation of typological frequency, and some readers may be worried about the effects of this choice. As far as we know, PHOIBLE is the most extensive archive of phonological inventories, so it is a reasonable choice in the current context. However, one may want to calculate typological frequency not by trying to have as many languages represented as possible, but rather by selecting a sample of typologically independent languages. In addition, it is not the case that all the world's languages are represented, and indeed some of the Yélî sounds were not found in PHOIBLE. PHOIBLE—as well as our own work—depends on phonological descriptions from linguists who are in many cases not native speakers of the languages. Because the phones in our items have largely been evidenced as phonemic via multiple analyses (i.e., minimal contrast, phonological, phonetic, and ultrasound, see Levinson, 2021), we are not concerned that changes to the phonolog-

ical description in the future (e.g., if a segment loses its phonemic status) will significantly change the results presented here. Relatedly, any converging evidence from the other ongoing studies of Yélî Dnye phonological development and fine-grained analyses of sound substitutions would certainly help bolster the claims we made here. While all these limitations should be borne in mind, it is important to also consider what our conclusions were, and that is that there is a non-trivial correlation between NWR and typological frequency. At present, we do not see how imbalance in the typological selection and missing data can conspire to produce the correlation we observe. If anything, these factors should increase noise in the typological frequency estimation, in which case the correlation size we uncover is an underestimation of the true correlation.

Additionally, we only had a single person interacting with children as well as interpreting children's production, so we do not know to what extent our findings generalize to other experimenters and research assistants. Furthermore, since both stimuli presentation and production data collected were audio-only, neither the children nor our research assistant were able to integrate visual production cues in their interpretation. Other work shows that children's performance reaches ceiling by 12 years of age for auditorily-presented minimal pairs for typologically rare (i.e., pre- vs post-alveolar stop) contrasts (Casillas & Levinson, In preparation). Nonetheless, language processing for the majority of children will be audiovisual in natural conditions, and thus it may be interesting in the future to capture this aspect of speech.

Conclusions

The present study shows that NWR can be adapted for very different populations than have previously been tested. In addition, we observed strong correlations with age and typological frequency, while correlations with item length, participant sex, maternal education, and birth order were weaker. A consideration of previous work led us to suggest that the statistical strength of all of these effects may vary depending on the linguistic, cultural, and socio-demographic properties of the population under study, in conjunction with characteristics of the non-word items used. The present findings raise many questions, including: Why do NWR scores pattern differently across samples? What does that tell us about the relationship between lexical development, phonological development, and the input environment? What is implied about the joint applicability of these outcome measures as a diagnostic indicator for language delays and disorders? While answers to these questions should be sought in future work, we take the present findings as robustly supporting the idea that phonological development continues well past early childhood and as yielding preliminary support for a potential association between individual learners' NWR and much broader patterns of cross-linguistic phone frequency.

References

Armon-Lotem, S., Jong, J. de, & Meir, N. (2015). *Methods for assessing multilingual*

children: Disentangling bilingualism from specific language impairment. Bristol: Multilingual matters.

Balladares, J., Marshall, C., & Griffiths, Y. (2016). Socio-economic status affects sentence repetition, but not non-word repetition, in Chilean preschoolers. *First Language, 36*(3), 338–351. https://doi.org/10.1177/0142723715626067

Barclay, K. J. (2015). A within-family analysis of birth order and intelligence using population conscription data on Swedish men. *Intelligence, 49*, 134–143.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language, 85*(4), 785–832.

Bermúdez-Otero, R. (2015). Amphichronic explanation and the life cycle of phonological processes. In P. Honeybone & J. Salmons (Eds.*), The Oxford handbook of historical phonology* (pp. 374–399). Oxford, UK: Oxford University Press.

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* (Version 6.1.35). Retrieved from http://www.praat.org/

Bowey, J. A. (2001). Nonword repetition and young children's receptive vocabulary: A longitudinal study. *Applied Psycholinguistics, 22*(3), 441–469.

Brandeker, M., & Thordardottir, E. (2015). Language exposure in bilingual toddlers: Performance on nonword repetition and lexical tasks. *American Journal of Speech-Language Pathology, 24*(2), 126–138.

Brown, P. (2011). The cultural organization of attention. In A. Duranti, E. Ochs, & and Bambi B Schieffelin (Eds.), *Handbook of Language Socialization* (pp. 29–55). Malden, MA: Wiley-Blackwell.

Brown, P. (2014). The interactional context of language learning in Tzeltal. In I. Arnon, M. Casillas, C. Kurumada, & B. Estigarribia (Eds.), *Language in interaction: Studies in honor of Eve V. Clark* (pp. 51--82). Amsterdam, NL: John Benjamins.

Brown, P., & Casillas, M. (in press). Childrearing through social interaction on Rossel Island, PNG. In A. J. Fentiman & M. Goody (Eds.), *Esther Goody revisited: Exploring the legacy of an original inter-disciplinarian* (pp. XX–XX). New York, NY: Berghahn.

Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Alam, F., … Casillas, M. (under review). A cross-cultural examination of young children's everyday language experiences. Available from https://psyarxiv.com/723pr/

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

Casillas, M., & Levinson, S. C. (In preparation). Markedness and minimal pair discrimination in children learning Yélî Dnye.

Castro-Caldas, A., Petersson, K. M., Reis, A., Stone-Elander, S., & Ingvar, M. (1998). The illiterate brain. Learning to read and write during childhood influences the functional organization of the adult brain. *Brain: A Journal of Neurology, 121*(6), 1053–1063. https://doi.org/10.1093/brain/121.6.1053

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from specific language impairment* (pp. 125–150). Bristol: Multilingual matters.

Chiat, S., & Roy, P. (2007). The preschool repetition test: An evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research, 50*(2), 429–443.

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language & Communication Disorders, 43*(1), 1–40.

COST Action. (2009). *Language impairment in a multilingual society: Linguistic patterns and the road to assessment.* Brussels: COST Office. Available Online at: http://www.bi-Sli.org.

Cristia, A., & Casillas, M. (2021). *Supplementary materials to non-word repetition in children learning Yélî Dnye.* Retrieved from https://osf.io/5qspb/wiki/home/

Cristia, A., Farabolini, G., Scaff, C., Havron, N., & Stieglitz, J. (2020). Infant-directed input and literacy effects on phonological processing: Non-word repetition scores among the Tsimane'. *PLoS ONE, 15*(9), e0237702. https://doi.org/10.1371/journal.pone.0237702

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research, 47*, 421–436.

Estes, K. G., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*, 177–195.

Farabolini, G., Rinaldi, P., Caselli, C., & Cristia, A. (2021). Non-word repetition in bilingual children: The role of language exposure, vocabulary scores and environmental factors. *Speech Language and Hearing*, 1–16. https://doi.org/10.1080/2050571X.2021.1879609

Farabolini, G., Taboh, A., Ceravolo, M. G., & Guerra, F. (2021). The association between language exposure and non-word repetition performance in bilingual children: A meta-analysis. Manuscript under review.

Farmani, H., Sayyahi, F., Soleymani, Z., Labbaf, F. Z., Talebi, E., & Shourvazi, Z. (2018). Normalization of the non-word repetition test in Farsi-speaking children. *Journal of Modern Rehabilitation, 12*(4), 217–224.

Foley, W. A. (1986). *The Papuan languages of New Guinea*. Cambridge, UK: Cambridge University Press.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language, 44*(3), 677–694.

Gallagher, G. (2014). An identity bias in phonotactics: Evidence from Cochabamba Quechua. *Laboratory Phonology, 5*(3), 337–378. https://doi.org/10.1515/lp-2014-0012

Gathercole, S. E., Willis, C., & Baddeley, A. D. (1991). Differentiating phonological memory and awareness of rhyme: Reading and vocabulary development in children. *British Journal of Psychology, 82*(3), 387–406.

Grätz, M. (2018). Competition in the family: Inequality between siblings and the intergenerational transmission of educational advantage. *Sociological Science, 5*, 246–269.

Havron, N., Ramus, F., Heude, B., Forhan, A., Cristia, A., Peyre, H., & Group, E. M.-C. C. S. (2019). The effect of older siblings on language development as a function of age difference and sex. *Psychological Science, 30*(9), 1333–1343.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics, 28*(4), 377–396.

Hellwig, B., Sarvasy, H., & Casillas, M. (provisionally accepted). Language acquisition. In N. Evans & S. Fedden (Eds.), *The Oxford guide to Papuan languages* (pp. XX–XX). Oxford: Oxford University Press.

Jaber-Awida, A. (2018). Experiment in non word repetition by monolingual Arabic preschoolers. *Athens Journal of Philology, 5*, 317–334. https://doi.org/10.30958/ajp.5-4-4

Kalnak, N., Peyrard-Janvid, M., Forssberg, H., & Sahlén, B. (2014). Nonword repetition–a clinical marker for specific language impairment in Swedish associated with parents 'language-related problems. *PloS One, 9*(2), e89544.

Kolinsky, R., Leite, I., Carvalho, C., Franco, A., & Morais, J. (2018). Completely illiterate adults can learn to decode in 3 months. *Reading and Writing, 31*(3), 649–677. https://doi.org/10.1007/s11145-017-9804-7

Lancy, D. F. (2015). *The anthropology of childhood*. Cambridge, UK: Cambridge University Press.

Lehmann, J.-Y. K., Nuevo-Chiquero, A., & Vidal-Fernandez, M. (2018). The early origins of birth order differences in children's outcomes and parental behavior. *Journal of Human Resources, 53*(1), 123–156.

Levinson, S. C. (2021). *A grammar of Yélî Dnye, the Papuan language of Rossel Island*. Berlin, Boston: De Gruyter Mouton.

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & de Vos, C. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science, 36*(4), 698–713. https://doi.org/10.1111/j.1551-6709.2011.01228.x

Maddieson, I. (2005). Correlating phonological complexity: Data and validation. UC *Berkeley PhonLab Annual Report, 1*(1).

Maddieson, I. (2009). Phonology, naturalness and universals. *Poznán Studies in Contemporary Linguistics, 45*(1), 131–140.

Maddieson, I. (2013a). *Consonant inventories. The World Atlas of Language Structures Online.* Retrieved from https://wals.info/chapter/1

Maddieson, I. (2013b). *Vowel quality inventories. The World Atlas of Language Structures Online.* Retrieved from https://wals.info/chapter/2

Maddieson, I., & Levinson, S. C. (in preparation). The phonetics of Yélî Dnye, the language of Rossel Island.

Meir, N., & Armon-Lotem, S. (2017). Independent and combined effects of socioeconomic status (SES) and bilingualism on children's vocabulary and verbal short-term memory. *Frontiers in Psychology, 8*, 1442.

Meir, N., Walters, J., & Armon-Lotem, S. (2016). Disentangling SLI and bilingualism using sentence repetition tasks: The impact of L1 and L2 properties. *International Journal of Bilingualism, 20*(4), 421–452.

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. Retrieved from https://phoible.org/

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning, part II: substance. *Language and Linguistics Compass, 6*(11), 702–718.

Ohala, J. J. (1981). The listener as a source of sound change. In M. F. Miller, C. S. Masek, & R. A. Hendrick (Eds.), *Papers from the parasession on language and behavior* (pp. 178–203). Chicago, IL: Chicago Linguistics Society.

Peute, A. A. K., Fikkert, P., & Casillas, M. (In preparation). Early consonant production in Yélî Dnye and Tseltal.

Piazzalunga, S., Previtali, L., Pozzoli, R., Scarponi, L., & Schindler, A. (2019). An articulatory-based disyllabic and trisyllabic Non-Word Repetition test: reliability and validity in Italian 3-to 7-year-old children. *Clinical Linguistics & Phonetics, 33*(5), 437–456.

Rumsey, A. (2017). Dependency and relative determination in language acquisition: The case of Ku Waru. In N. J. Enfield (Ed.), *Dependencies in language* (pp. 97–116). Berlin: Language Science Press.

Santos, C. dos, Frau, S., Labrevoit, S., & Zebib, R. (2020). L'épreuve de répétition de non-mots LITMUS-NWR-FR évalue-t-elle la phonologie? *Congrès Mondial de Linguistique Française, 78*, 10005.

Scaff, C. (2019). Beyond WEIRD: An interdisciplinary approach to language acquisition (PhD thesis).

Scaff, C., Stieglitz, J., Casillas, M., & Cristia, A. (under review). Daylong audio recordings of young children in a forager-farmer society show low levels of verbal input with minimal age-related changes.

Stokes, S. F., Wong, A. M., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research, 49*, 219–236.

Torrington Eaton, C., Newman, R. S., Ratner, N. B., & Rowe, M. L. (2015). Non-word repetition in 2-year-olds: Replication of an adapted paradigm and a useful methodological extension. *Clinical Linguistics & Phonetics, 29*(7), 523–535.

Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., … Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany. International *Journal of Language & Communication Disorders, 53*(4), 888–904.

Vance, M., Stackhouse, J., & Wells, B. (2005). Speech-production skills in children aged 3–7 years. *International Journal of Language & Communication Disorders, 40*(1), 29–48.

Wilsenach, C. (2013). Phonological skills as predictor of reading success: An investi-

gation of emergent bilingual Northern Sotho/English learners. *Per Linguam: A Journal of Language Learning= Per Linguam: Tydskrif Vir Taalaanleer, 29*(2), 17–32. https://doi.org/10.5785/29-2-554

Yu, A. C. L. (2021). Toward an individual-difference perspective on phonologization. *Glossa: A Journal of General Linguistics, 6*(1), 1–24.

Acknowledgments

We are grateful to the individuals who participated in the study, and the families and communities that made it possible. The collection and annotation of these recordings was made possible by Ndapw:éé Yidika, Taakêmê Ńamono, and Y:aaw:aa Pikuwa; with thanks also to the PNG National Research Institute, and the Administration of Milne Bay Province. We owe big thanks also to Stephen C. Levinson for his invaluable advice and support and Shawn C. Tice for helpful discussion during data collection. AC acknowledges financial and institutional support from Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017) and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. MC acknowledges financial support from an NWO Veni Innovational Scheme grant (275-89-033).

Data, Code and Materials Availability Statement

All data, code, and materials are available from https://osf.io/5qspb/

Ethics statement

This study was approved as part of a larger research effort by the second author. The line of research was evaluated by the Radboud University Faculty of Social Sciences Ethics Committee (Ethiek Commissie van de faculteit der Sociale Wetenschappen; ECSW) in Nijmegen, The Netherlands (original request: ECSW2017-3001-474 Manko-Rowland; amendment: ECSW-2018-041), including the use of verbal (not written) consent. Child assent is also culturally pertinent. See main text, Participants section, for further information.

Authorship and Contributorship Statement

Both authors contributed to study funding, design, data collection, annotation, analyses, writing.

License

A demonstration of the uncomputability of parametric models of language acquisition and a biologically plausible alternative

Evelina Leivada
Universitat Rovira i Virgili, Spain

Elliot Murphy
University of Texas Health Science Center at Houston, USA

Abstract: The logical problem of language acquisition has been at the forefront of psycholinguistics and behavioral neuroscience for decades. One of the most influential answers to the problem of how successful acquisition occurs on the basis of noisy input suggests that the child is aided by innate principles and parameters (P&P). These are conceived as part of our biological endowment for language. Previous work on the computability of parametric models has focused on the process of parameter-setting, leaving settability unaddressed. Settability is a key notion in parametric models since it provides an answer to the logical problem of language acquisition: the setting of one parameter carries implications for the settability of others, minimizing the child's task. However, a mathematical analysis of the expected probability of successful computation of settability relations has not been carried out. We report results from a novel program developed to calculate the probability of successful computation of a network of 62 linguistic parameters as attested in 28 languages, spanning across 5 language families. The results reveal that some parameters have an extremely low probability of successful computation, such that trillions of unsuccessful computations are expected before a successful setting occurs. Using the same program, we performed an additional analysis on a different network, covering 94 parameters from 58 languages and 15 language families. In this case, the estimated number of expected unsuccessful computations rose from trillions to quadrillions. These results raise concerns about the computational feasibility of the highly influential P&P approach to language development. Merging insights from various acquisition models, including some developed within P&P, a biologically plausible alternative is offered for the process of deciphering a target grammar in the acquisition of both spoken and signed languages. Overall, our analysis of the P&P approach to language acquisition centers learnability and computability constraints as the major factors for determining the psychological plausibility of grammar development.

Keywords: acquisition; computation; learnability; parameter; Universal Grammar

Corresponding author(s): Evelina Leivada, Department of English and German Studies, Universitat Rovira i Virgili, Av. Catalunya 35, 43002, Tarragona, Spain. Email: evelina.leivada@urv.cat

ORCID ID(s): https://orcid.org/0000-0003-3181-1917, https://orcid.org/0000-0003-1456-0343

Citation: Leivada, E., & Murphy, E. (2022). A demonstration of the uncomputability of parametric models of language acquisition and a biologically plausible alternative. *Language Development Research, 2*(1), 105–138. https://doi.org/10.34842/2022-585

Introduction

Language acquisition and its guiding principles have been at the forefront of psycholinguistic and developmental research for over five decades. Among the central research questions of the field, the logical problem of language acquisition stands out: How is language acquired, given the noisy nature of the linguistic input that a child receives during the early stages of development? The poverty of the environmental stimulus that characterizes the input sharply contrasts with the richness of the attainment that a neurotypical child will have as a mature speaker/signer (Chomsky, 1965; 1980). In (bio)linguistics and psychology, the highly influential Principles & Parameters framework (P&P) has provided an answer to the logical problem of language acquisition by positing that the child is aided by some innate principles that help them navigate the space of cross-linguistic variation in the process of acquisition (Chomsky, 1981). According to P&P, the child is innately equipped with a cognitive apparatus called Universal Grammar. Universal Grammar can be viewed as a cognitive map that consists of (i) a finite number of universal principles and (ii) a small number of parameters, that are also universal, but come with a set of values to which they are variably set across different languages. Although this idea has been around for several decades and has been criticized on various grounds, recently there has been a renewed interest in it, especially from a computational perspective that integrates Universal Grammar and non-linguistic principles of computation in the process of language development (Yang et al., 2017; Kazakov et al., 2018; Manzini, 2019). Even though the P&P framework removes some of the burden originally placed on the Evaluation Measure, it remains unclear what type of learning algorithm can manoeuvre itself through a space of grammars.

From a theoretical perspective, this organization of Universal Grammar in terms of principles and parameters brings an important benefit. Consider the overall volume of the input data a child has to process in order to acquire their language. Not only is it vast, but the task at hand entails dealing with noisy data and complex rules, whose properties the child has to decipher in the earliest stages of development. The logical problem of language acquisition addresses the question of how the child achieves this monumental task. The answer, within the P&P framework, is that the child's cognitive map consists of a finite number of parameters that form certain paths (Figure 1), such that the variation space is neatly compartmentalized, rendering the child's task considerably easier.

To explain the process, at point zero of acquisition the child has routes of the cognitive map open, but upon setting a few initial parameters to one value instead of another, the child selects a path. This selection brings with it the notion of *settability*: The values of the first-set parameters carry implications about the settability of others that are yet to be set. After selecting a route through setting a parameter to one value, the child is bound against exploring other routes, at least not in the context of that language. Parameters in these other routes will not be set to a value on the basis of the data the child is exposed to, because they are not settable: they do not form part of the route the child has taken. Since the child will never have to deal with them, the

variation space that they have to navigate is substantially reduced. This explains (putatively) how the child performs this complex task so fast.

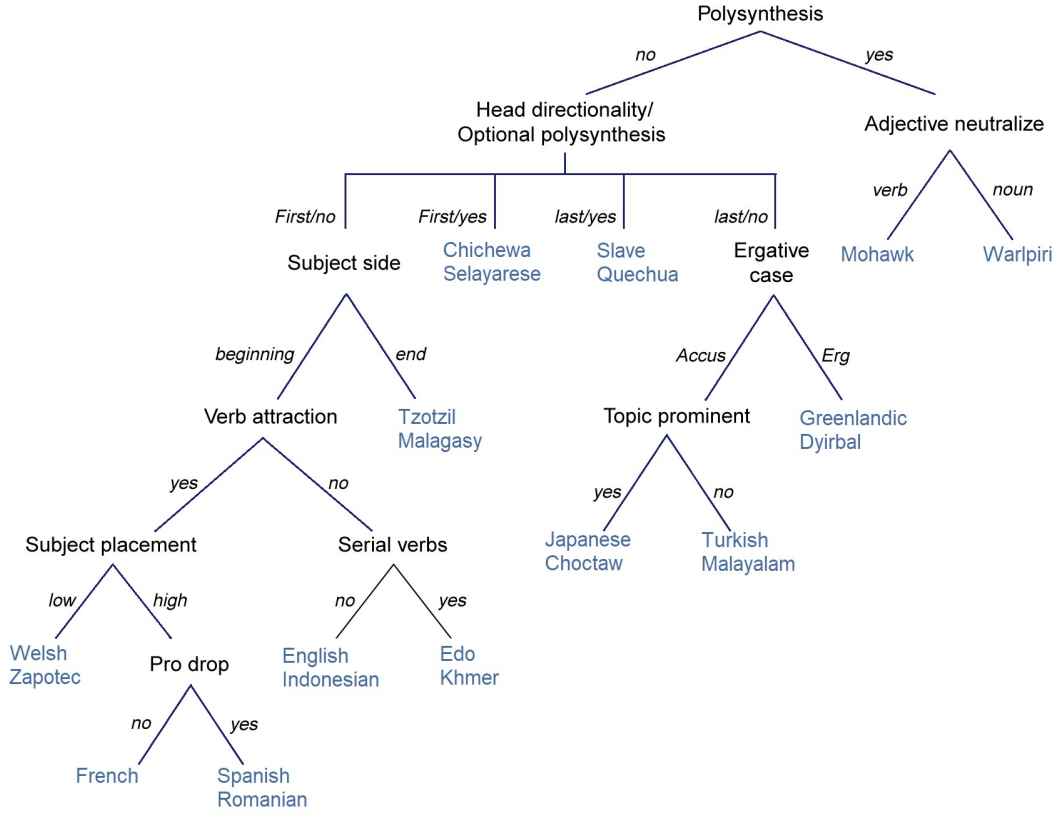


Figure 1. ***A parametric hierarchy (adapted from Baker 2003).***

To define the critical notions of setting and settability, parameter setting refers to selecting a value for a parameter, based on data from the target language. Parameter settability refers to whether a parameter forms part of the route the learner has taken. To give an example based on Figure 1, 'adjective neutralize' is not settable in French (i.e., it does not form part of the route to French), but it is settable and set to a specific value in Mohawk. A settable parameter is always set based on language data. Therefore, setting differs from settability in that the latter only arises given the existence of an *implicational network* among parameters (i.e., a network of dependencies that specifies that the settability of parameter X depends on having set parameter Y to one value instead of another, as shown in Figure 1). In this sense, the crucial difference between the two notions, setting and settability, boils down to the fact that the process of setting/value selection does not bear upon the existence of an implicational network; the latter is only informative about settability.

The processes of setting and settability are formally presented in (1).

(1a) Setting

Given a parametric hierarchy composed by a series of nodes N_i, where $i \in \mathbb{N}$, setting is the process of selecting a binary value $S_i = \{+N_i \; or \; -N_i\}$. If you get input z, select a value. If z matches the hypothesized value, set N to this value. If not, select the other value and set N. Reach state N_{valued}.

(1b) Settability

Go to the next node N2. Check whether there is a path that connects N2 to any previous node (in this example, N_{valued}). A path entails a logical expression (e.g., N2=(N−)). If there is no path, set N2 following the process described in *Setting*. If there is a path, determine its satisfiability. A path is satisfied if the parts of the logical expression match the values of previously set nodes (e.g., if the logical expression is N2=(N−), then N_{valued} must be set to −. If it is not, the path is not satisfied and settability of N2 cannot be reached on this path). Repeat for every path that connects N2 to previous nodes. If one (or more than one, but at least one) path is satisfied, follow the process described in *Setting* to set N2. If no path is satisfied, rewrite N2 as $N2_{not\text{-}settable}$.

There are two possible outcomes: $N2_{valued}$ or $N2_{not\text{-}settable}$. Once any of the two is reached, go to the next node N3 and repeat the process. When all nodes have reached one of the two states, N_{valued} or $N_{not\text{-}settable}$, halt the process.

These two notions, setting and settability, have not been investigated to equal degrees. Previous work concerning the computation of parametric models of language acquisition has focused almost exclusively on analyzing setting relations; for example, the number of linguistic examples and initial hypotheses that are needed for the child to set the parameters that correspond to their target language (Gibson & Wexler, 1994; Niyogi & Berwick, 1996). Settability has not been addressed from a computational perspective, in part because until recently it was largely assumed that there is only one way of reaching settability for a given parameter in a given language; an assumption that voids the need for further computation.

To illustrate this assumption, Figure 1 shows that there is a single way to reach the settability of any parameter in this parametric hierarchy (e.g., 'adjective neutralize' is reached exclusively by setting polysynthesis to [+]). The only work that addresses the computation of settability relations challenged this assumption of unique settability (Boeckx & Leivada, 2013), through examining an elaborate network of parameters from the nominal domain (henceforth, the network, Longobardi & Guardiano, 2009; Figure 2).

No	Parameters & Paths	It	Sal	Sp	Fr	Ptg	Ru	Lat	ClG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Ba
1	± grammatical person	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2	± grammatical number (+1)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3	± grammatical gender (+2)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
4	± variable person on D (+1)	-	-	+	-	-	-	?	+	?	-	+	?	?	-	-	-	-	-	-	?	?	-	+	-	-	+
5	± feature spread to N (+2)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-
6	± number on N (Bare Nouns) (+5)	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	0
7	± grammaticalized partial definiteness	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	-
8	± grammaticalized definiteness (+7)	+	+	+	+	+	+	0	+	+	+	+	-	+	+	+	+	+	0	0	+	+	+	+	+	+	0
9	± free null partitive Q (+6)	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
10	± grammaticalized distal article (-5 or -6 or +7)	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	-	+	-	+
11	± grammaticalized topic article (-10)	-	+	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	-	0	-	0
12	± definiteness checking N (+7)	-	-	-	-	-	+	0	-	-	-	-	-	-	-	-	+	+	0	0	-	-	-	-	-	-	0
13	± definiteness spread to N (+12)	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	+	-	0	0	0	0	0	0	0	0	0
14	± definiteness on attributes (+7, -12)	-	-	-	-	-	0	0	+	+	-	+	+	-	-	-	0	0	0	0	-	-	+	+	-	-	0
15	± definiteness on relatives (+7)	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	+	+	-	0
16	± D-controlled inflection on N (+5)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	0	-	0
17	± grammaticalized cardinal nouns	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	?	+	+	?	+	+	+	?	-	-
18	± grammaticalized cardinal adjectives (+17)	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	?	-	+	?	-	-	+	?	0	0
19	± plural spread from cardinals (+5, -17 or +18)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	0	+	?	0	0	+	0	-	0
20	± grammaticalized mass-to-count	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	?	+	-	+	-	-	-
21	± N-to-predicate incorporation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-
22	± grammaticalized partial count (-5 or -6 or +7,-21)	+	+	+	+	+	+	0	-	-	+	+	0	+	+	+	+	+	0	0	-	-	-	-	0	0	+
23	± grammaticalized count (+22)	+	+	+	+	+	+	0	0	0	+	+	0	-	+	+	+	-	0	0	0	0	0	0	0	0	-
24	± count-checking N (+21 or +22)	-	-	-	-	-	-	0	0	0	-	-	0	-	-	-	-	-	0	0	0	0	0	0	-	-	-
25	± prepositional Genitive	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	+	+	+	-	-	-
26	± free inflected Genitive (-25)	0	0	0	0	0	-	+	+	-	-	-	-	-	0	0	0	0	-	-	0	0	0	0	-	-	-
27	± Genitive O (+25 or -26)	-	-	-	-	-	-	0	0	+	+	+	+	+	-	+	?	-	+	+	+	+	-	-	+	-	+
28	± Genitive S (+25 or -26)	-	-	-	-	-	+	0	0	-	-	-	+	+	+	+	+	-	-	-	-	-	+	+	-	+	+
29	± postpositional Genitive (+27 or +28)	0	0	0	0	0	-	0	0	-	-	-	-	-	+	-	+	0	-	-	-	-	-	-	-	-	+
30	± Genitive over DemP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	?
31	± poss-checking N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-
32	± structured APs	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-
33	± feature spread to structured Aps (+32)	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	?	+	+	0	-	-
34	± feature spread to predicative APs	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	?	+	+	+	+	+
35	± number on A (+6,+33 or +34)	+	+	+	0	+	+	+	+	+	+	+	+	+	0	+	+	+	+	+	+	?	+	+	0	+	0
36	± D-controlled inflection on Adjectives (+33)	-	-	-	-	-	-	-	-	-	-	-	+	+	0	+	+	-	-	-	-	?	-	-	0	0	0
37	± DemP over relative clauses	+	+	+	+	+	+	?	?	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
38	± free APs in Modifier Phrase (+32)	+	+	+	+	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	+	0	-	-
39	± APs in Modifier Phrase (-38)	0	0	0	0	0	-	0	0	0	-	0	0	+	+	-	-	+	+	+	-	-	+	0	0	-	?
40	± overt Mod° (-32 or +38 or +39)	-	+	-	-	-	0	-	-	-	0	-	-	-	-	0	0	-	-	-	0	0	-	-	+	0	?
41	± adjectival Genitive	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-
42	± N-raising with pied-piping	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+
43	± N over external argument (-42)	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	0	0	+	+	0
44	± N over Genitive O (+26 or +27, -30, +43)	0	0	0	0	0	0	-	-	+	+	+	+	+	0	+	0	0	+	+	+	+	0	0	+	0	0
45	± N over Adjectives (+32, (-26 or -27, +43) or +44)	+	+	+	+	+	+	0	0	-	+	-	-	-	0	-	0	-	-	-	+	+	0	0	0	-	0
46	± N over Manner 2 Adjectives (+45)	+	+	+	+	+	+	0	0	0	+	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0
47	± N over Manner 1 Adjectives (+46)	-	+	-	-	-	-	0	0	0	?	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0
48	± N over high Adjectives (+47)	0	+	0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0	0	-	-	0	0	0	0	0
49	± N over cardinals (+42 or +48)	0	-	0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0	0	0	0	+	+	?	0	-
50	± strong D (person) (+1, +8 or +28)	+	+	+	+	+	+	0	+	+	+	+	0	-	-	-	-	+	0	0	-	-	+	+	-	?	+
51	± NP over D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+
52	± N strong deixis	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+
53	± strong anaphoricity (+52)	+	+	-	+	+	-	-	-	-	+	-	-	+	+	+	+	-	+	+	0	0	?	-	+	+	+
54	± DP over Demonstratives (-51, +52)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	+	?	0	-	0
55	± D-checking Demonstratives (-5 or -6 or +7, +52)	+	+	+	+	+	+	0	-	-	+	-	?	+	+	+	-	+	0	0	0	0	+	-	+	-	+
56	± D-checking possessives (-5 or -6 or +8, +50 or -28)	-	-	+	+	?	-	0	-	-	-	-	0	0	0	0	0	-	0	0	+	+	-	-	+	?	-
57	± feature spread on possessives (+33 or +34 or +35)	+	+	+	+	+	+	+	+	+	-	-	+	+	0	+	+	+	0	0	-	?	-	-	+	-	-
58	± feature spread on postpositional Genitive (+29, +57)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0
59	± enclitic possessives	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	?	-	-	-	-	-	-	?	-	-
60	± Consistency Principle (-43 or -44 or -45 or -46 or -47 or +51)	+	0	+	+	+	+	?	?	?	?	-	?	+	+	+	+	-	-	-	0	0	0	0	?	+	+
61	± null N-licensing article (-5 or -6 or -12, +50 or +51)	-	-	+	-	+	0	0	+	+	-	-	0	0	0	0	0	0	0	0	0	0	-	-	?	?	+
62	± obl. def. inheritance (+7, -22, (-25, +26) or +27, +42 or +45 or -50)	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	-	+	+	+	-	0	0
63	± gramm. geographical article (-5 or -6 or +7,-22 or -23 or +45)	+	+	-	+	-	+	0	+	+	+	0	0	0	0	0	0	-	0	0	?	?	-	-	0	-	-
		It	Sal	Sp	Fr	Ptg	Ru	Lat	ClG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Ba

Figure 2. ***The analyzed network consists of 63 binary parameters from the nominal domain across 28 languages (adapted from Longobardi & Guardiano, 2009). The first column presents the parameters and the settability path(s) on which each parameter is settable. If a settability path is not available in a language, the corresponding parameter is marked with 0 (e.g., if [5settable] depends on [4−], if the latter is in any other state, the former is marked with 0, which indicates that the parameter is not settable in the specific language). ',' means ∧.***

The assumption of unique settability was investigated through the use of a program that calculated whether the settability paths in Figure 2 were satisfied in each language-parameter pairing that exists in the network (Boeckx & Leivada, 2013). For example, if the network specifies that the settability of parameter (P) 14 is reached on

the basis of setting P7 to + and P12 to −, the settability path would be: [14settable] = [7+] AND [12−]. The program read these paths in the form of logical expressions and checked whether they were satisfied in the input it received. The input was the states of each parameter in each language, as they are shown in the language columns of Figure 2. Proceeding with the previous example, if in language X, P7 was set to + and P12 was set to −, the program returned the outcome 'true' for [14settable]. If P7 and P12 were in any other state (i.e., set in the opposite value or not-settable), the program returned the outcome 'false', which means that P14 is not settable (on this settability path) for language X.

As the 'OR' nodes in the first column of Figure 2 suggest, the network makes available different paths for the settability of many of its parameters. Until the computation of the settability relations of every language-parameter pairing, it was unclear whether different settability paths existed for different languages or whether the same language could involve more than one path for the same parameter; something that would disprove the assumption of unique settability. Previous work on the computation of settability relations determined that there are different ways to reach settability of a parameter, not only across but also within languages (Boeckx & Leivada, 2013). For example, Table 1 shows that for many languages in the network, parameter 29 is not settable (e.g., It[alian] in the second column). For other languages, the parameter is settable in one (e.g., path 4 in Rom[anian]) or more ways (e.g., paths 3 and 4 in Ba[sque])

Table 1. ***Parameter 29: ± Postpositional Genitive. 1 signals the availability of the corresponding settability path in the relevant language, whereas 0 signals the unavailability of the path. When a number node in the first column has an attached parenthesis on its right (e.g., 2+(1+)), the node inside the parenthesis is the settability path of the node outside the parenthesis, until an independent parameter is reached. In this table, the settability of parameter 29 is possible on the basis of setting either 27 to + or 28 to +. Both 27 and 28 are dependent parameters, settable in two ways each, either through setting 25 to + or 26 to −. Parameter 25 is an independent parameter which means that its settability does not depend on the setting of other parameters (Boeckx & Leivada, 2013).***

4 Paths	**It**	**Sal**	**Sp**	**Fr**	**Ptg**	**Rom**	**Lat**	**ClG**	**NTG**	**Gri**	**Grk**	**Got**	**OE**	**E**	**D**	**Nor**	**Blg**	**SC**	**Rus**	**Ir**	**Wel**	**Heb**	**Ar**	**Wo**	**Hu**	**Fin**	**Hi**	**Ba**
27+(25+)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
28+(25+)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0
27+(26-(25-))	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	1
28+(26-(25-))	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

However, there is a crucial and thus far unproven assumption behind previous work on the computation of settability relations. The program that was used in Boeckx &

Leivada (2013) was a semi-automatic one: the settability paths were not computed by it, but were given to it as predefined logical expressions. The crucial assumption is that the processing system, be it the human cognitive parser or a custom-made program that simulates the process of computation, *can* successfully compute more than one settability path for a single parameter. If the settability of a parameter can be determined in more than one way, the parser must engage in some kind of computation that exhaustively checks all the paths that lead to it in order to determine whether the parameter is settable. This happens because the process of determining settability is a necessary prerequisite for the process of parameter-setting. Given that (i) not all parameters are settable in all languages and (ii) the learner does not know a priori which parameters are not, because the settability paths become available progressively, depending on the value of earlier set parameters, the learner must engage in some kind of computation that determines whether the parameter that it encounters next in the hierarchy is settable or not.

The aim of the present work is to spell out the computation of settability relations, *locally* for each dependent parameter of the analyzed network. More specifically, by means of treating each settability path as a logical expression (examples (2)-(3)), the satisfiability of each path must be calculated by the parser, be it the human brain or, in this case, a program that will simulate the computational process. In terms of the parametric network that will be analyzed, the notion of satisfiability refers to whether a path involves parameter values that match the input (given in Figure 2), such that this path is available in a language-parameter pairing, making the parameter settable in the specific language.

(2) The logical expression for the second settability path of P10: $(5-) \wedge (2+) \wedge (1+)$

(3) The logical expression for all the paths of P10: $((5-) \wedge (2+) \wedge (1+)) \vee ((6-) \wedge (5+) \wedge (2+) \wedge (1+)) \vee (7+)$

The computation that follows operates on the basis of two important characteristics of the network and the learner respectively. First, if a parameter involves more than one settability path, the computation does not halt after finding a satisfiable path for a parameter. Instead, all paths need to be checked for satisfiability. In order to understand this characteristic, it is necessary to take into account that a parameter's settability paths often materialize at different times. For example, Table 2 shows that for P24, the first path becomes available after P21 is set to +, while the second path materializes after P22 is set to +. Even if the availability of a path for P24 was to be checked when [21+] was achieved, the computation would need to be re-run when [22+] was achieved, because not all languages set P24 on the first path (e.g., Ba in table 2).

Table 2. ***Parameter 24: ± Count-Checking N. 1 signals the availability of the corresponding settability path in the relevant language, whereas 0 signals the unavailability of the path. ',' means ∧ (Boeckx & Leivada, 2013).***

4 Paths	**It**	**Sal**	**Sp**	**Fr**	**Ptg**	**Rum**	**Lat**	**ClG**	**NTG**	**Gri**	**Grk**	**Got**	**OE**	**E**	**D**	**Nor**	**Blg**	**SC**	**Rus**	**Ir**	**Wel**	**Heb**	**Ar**	**Wo**	**Hu**	**Fin**	**Hi**	**Ba**
21+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
22+(7+, 21-)	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
22+((5-(2+(1+))), 21-)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22+((6-(5+(2+(1 +)))), 21-)	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The second important characteristic is that the learner cannot remember the parametric nodes that formed part of a previously checked path and reuse this information when checking paths that materialize later. For example, the last two paths of P24 in Table 2 share some parameters. Still, the satisfiability of the logical expression needs to be checked for the last path too. The reason has to do with memory limitations. Even setting aside interference concerns that arise from keeping track of paths that materialize *at different points* (hence are separated by the setting of other parameters that occurs in between their materialization), working memory has a capacity of maintaining four units, on average (Cowan, 2000). A set of three parametric nodes and their values already exceeds this capacity, and most paths are considerably longer than this.

Not only is this information not retainable in memory due to its heavy load, but the parser does not have memory that goes beyond the current state. Parametric models in language acquisition have long been described as involving memoryless processing, in the sense that at any step the learner has no recall of prior input or states, beyond the ones currently entertained (Page, 2004; Fodor & Sakas, 2005; Fodor, 2009). This memoryless character of the learning process has also been a crucial assumption in prior work on the computation of setting relations in parametric models (Niyogi & Berwick, 1996; 1997). We stress that while aspects of contemporary neuroscience support a view of the brain's memory as being capable of a pushdown stack (beyond deterministic pushdown automata), the mature state of a mildly context-sensitive grammar would presumably not be attainable immediately to the infant (Gallistel & King, 2009). This means that when the learner deals with the settability of P24 (Table 2), it cannot shorten its last three paths through rewriting P22 as settable/non-settable (i.e., it will not remember whether P22 was or was not settable in a language and replace the paths that determine its settability with this information), because it lacks the read/write memory of a Turing machine. Put differently, the four paths of P24 that are shown in Table 2 cannot be rewritten as two paths, 21+ and 22+, by means of

collapsing the different ways of reaching the settability of the latter. Additionally, such a move, apart from clashing with standard assumptions about properties of the brain, would raise empirical concerns. For example, Table 2 shows that French sets this parameter on two paths that depend on two different ways of reaching the settability of P22 (i.e., paths 2 and 4). Collapsing these two into one would simply not capture the facts for this language.

Taking into account these two characteristics, the present work aims to determine the computability of the settability paths behind the parameters of the analyzed network, through calculating the probability of running into loops that impede halting. For the computation to be successful, the learner needs to check the satisfiability of all the settability paths behind a parameter and halt. For example, if a parameter involves only two settability paths, A and B, further computation is not necessary, because the parser keeps track of the current state and will proceed to the next path without running into a loop: after checking both paths in one of the two possible orders, AB or BA, the computation will halt successfully. We can thus say that a parameter that has two settability paths has two ways of computation (i.e., two ways or orders of parsing the set of two paths): AB and BA. However, as the number of paths grows, the number of ways a set of paths can be checked for satisfiability also grows: two paths have two possible ways of computation (AB or BA), three paths have six ways (ABC, ACB, BAC, BCA, CAB, CBA), etc. In order to determine to what degree the ways of computation grow in the parametric network under examination, the program we describe below was designed to automatically calculate the probability of successful computation for each dependent parameter of the network, by estimating the ratio of successful computation to unsuccessful computation. The former refers to the number of ways the entire set of paths behind a parameter can be computed (i.e., checked for satisfiability) without running into loops; the latter refers to the number of ways that it runs into a single loop.

Method

The Longobardi & Guardiano network (Figure 2) consists of 63 parameters in 23 contemporary and 5 ancient languages, mostly from the Indo-European family. It is one of the most detailed parametric networks in the literature, rendering it an ideal candidate for computing settability relations. The present analysis used the slightly amended version of the network that was presented in previous work on the computability of parametric relations (Boeckx & Leivada 2013), in which parameter 62 was eliminated due to errors in its formulation. This elimination reduces the total number of the discussed parameters from 63 to 62. From these 62 parameters, 21 are settable on more than two paths, and hence these are the parameters analyzed in the present work.

In order to calculate the number of possible ways of successful computation (i.e., no loop), for n number of paths, $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$. For example, if n = 3, in the first random selection of a path, there are three options to choose from. In the second selection, there are n−1 options, and in the third selection, there are n−2

options, since no repetitions are permitted in successful computation. Therefore, for $n = 3$, there are $3 \cdot 2 \cdot 1 = 6$ ways of computation that do not run into a loop.

Calculating the ways of computation that feature a loop (i.e., one repetition of a previously checked path), in the first and the second random selection of a path, there cannot exist any repetition: in the first one, there is nothing to be repeated, and in the second one, the first selection will be remembered as the current path from which the learner is moving. From the third random selection of a path onwards, the total number of ways of unsuccessful computation is the sum of the different ways of unsuccessful computation when we take a subset k of n. The formula to calculate this is the following:

$$\sum_{k=3}^{k=n} \mathrm{n} \cdot (\mathrm{n}-1) \cdot \ldots \cdot \big(\mathrm{n}-(\mathrm{k}-2)\big) \cdot (\mathrm{k}-2)$$

For example, if a parameter has 5 settability paths ($n = 5$), for $k = 1$ and $k = 2$, there cannot be any repetitions. For $k = 3$, where 3 is the third random selection of a path, the number of ways of computation without repetition is $n \cdot (n-1) = 5 \cdot (5-1) = 20$. In order to calculate the number of ways of computation that feature a repetition in this third selection, this number must be multiplied by the number of paths that can be repeated. This is $k-2$ because the learner keeps track of the current state, so it cannot repeat the path it last checked. Therefore, for the third selection, $(k-2) \cdot 20$ gives a total of 20. For $k = 4$, the possible ways of computation without repetition are $n \cdot (n-1) \cdot (n-2) = 5 \cdot 4 \cdot 3 = 60$. This is multiplied by the number of paths that can be repeated, which is $k-2 = 2$, thus for $k = 4$, the number of ways of computation that have a repetition is $60 \cdot 2 = 120$. For $k = 5$, the possible ways of computation without repetition are $n \cdot (n-1) \cdot (n-2) \cdot (n-3) = 120$. This is multiplied by the number of paths that can be repeated, which is $k-2 = 3$, so for $k = 5$, the total number of ways of computation with repetition is $120 \cdot 3 = 360$. Overall, the total number of ways of unsuccessful computation for $n = 5$ is $360 + 120 + 20 = 500$.

For $n = 5$, the number of possible computations with and without loops is small, hence easy to calculate. However, many of the parameters in the analyzed network involve more than 10 paths. For this reason, a program was developed in Python in order to carry out the computation automatically (see Appendix for code). The program asks the user to provide the number of paths that should be computed. Upon being given a number followed by 'enter', it performs the calculation and asks the user whether they wish to perform another calculation for a different number of paths. Pressing '1' and then 'enter' restarts the process for another calculation, while pressing '2' and 'enter' closes the program. The program can be used to perform these calculations for any parametric model.

Results

The analysis produced two results: (i) the number of ways of successful and unsuccessful computation and (ii) the probability of successful computation for each parameter. Computation here refers not to the process of parameter-setting, but rather to going through the settability paths behind each parameter by means of checking the satisfiability of the logical expressions behind the paths ((2)-(3)). As noted, Figure 2 shows the Longobardi & Guardiano network. However, it provides no information as to how many paths the learner has to go through in order to determine settability and how many ways of computation (i.e., the process of "going through the set of paths") exist. Figure 3 addresses this gap by showing the degree to which the numbers for successful and unsuccessful computation rise in relation to the number of paths. More specifically, the average number of paths for the analyzed parameters is 8. For $n = 8$, there are 40,320 ways of successful computation and 375,368 ways of unsuccessful computation. This means that when a parameter has 8 settability paths, the memoryless parsing process has a total of 415,688 ways of going through them in order to check their satisfiability. For 10 paths, the number rises to 3,628,800 ways of successful computation and 46,253,610 ways of unsuccessful computation, while for 12 paths, the equivalent numbers are 6,227,020,800 and 1.11471e+11.

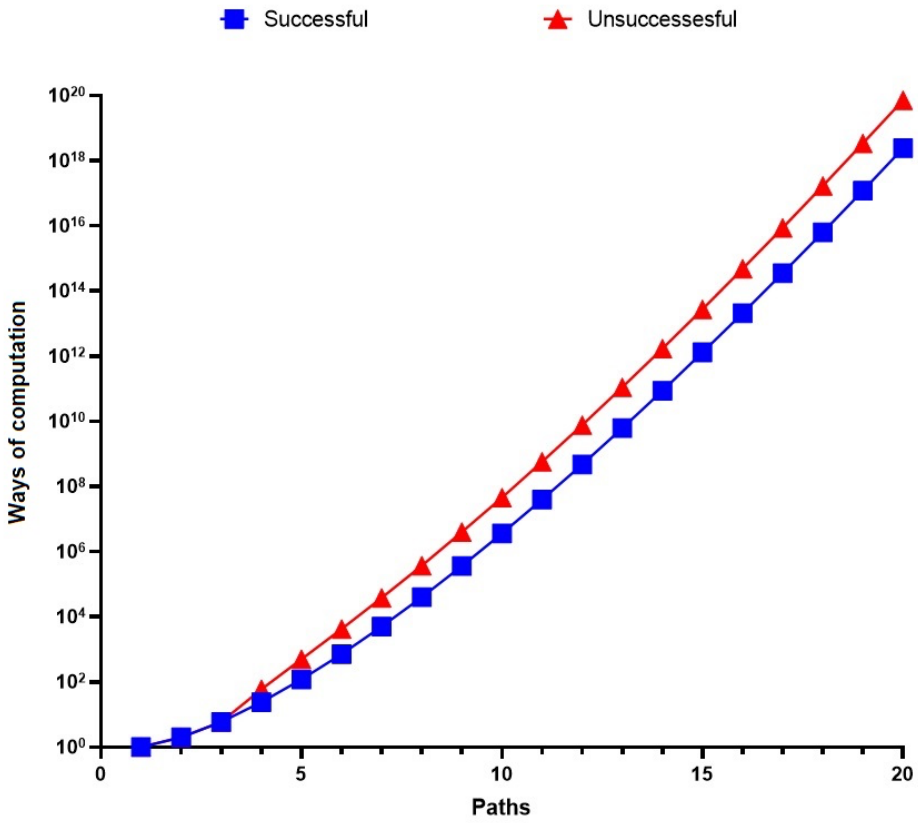


Figure 3. ***Number of ways of successful and unsuccessful computation across paths.***

Focusing on the Longobardi & Guardiano network, Figure 4 shows the ways of successful and unsuccessful computation for the parameters that have 3 or more settability paths. With the exception of the parameters that have just 3 paths, for all other parameters, the number of computations that run into a loop is considerably higher than the number of successful computations.

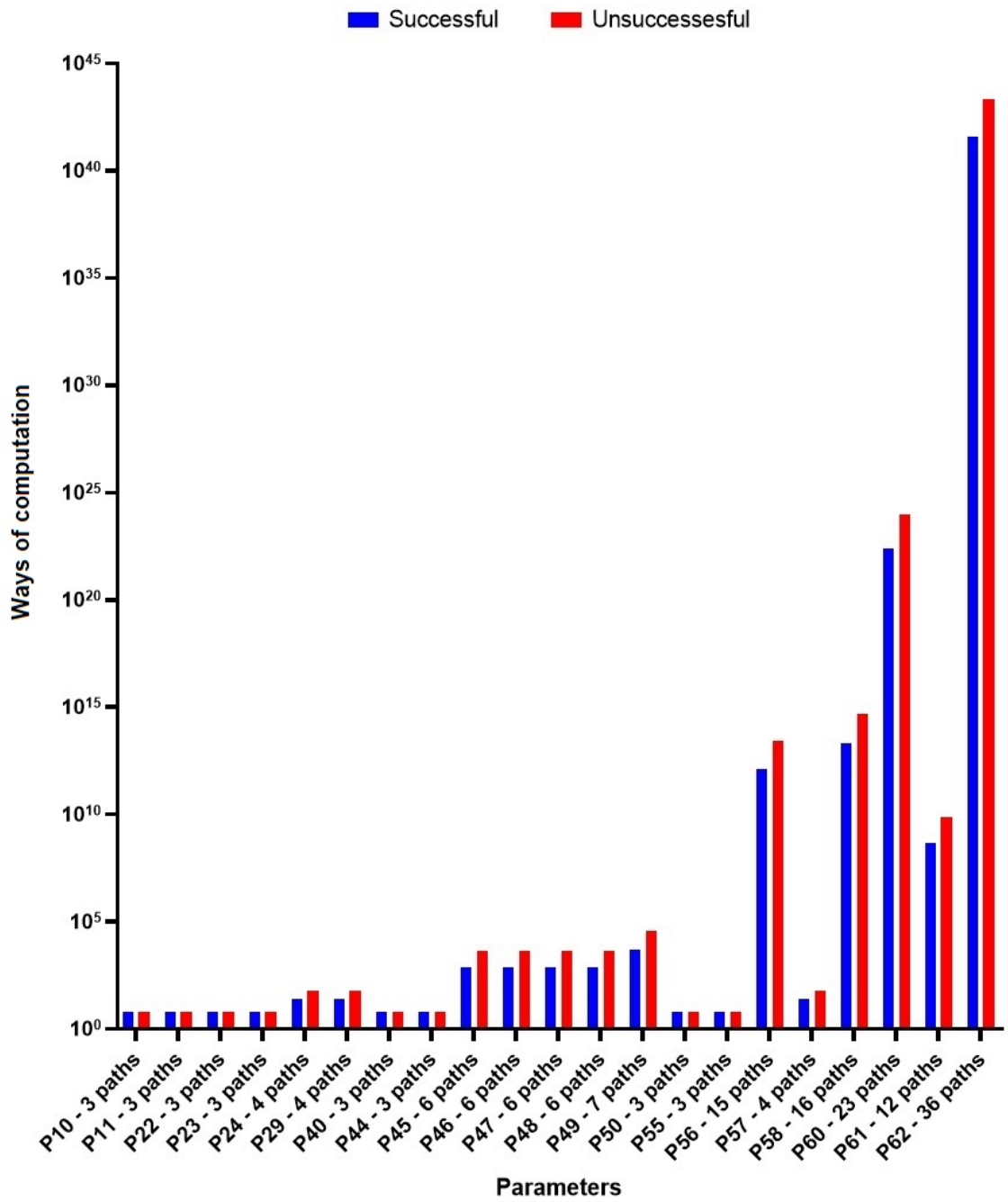


Figure 4. ***Number of ways of successful and unsuccessful computation for the 21 parameters of the analyzed network.***

The analyzed parameters involve a total of 169 settability paths. The probability of successful computation for each parameter independently is given in Figure 5.

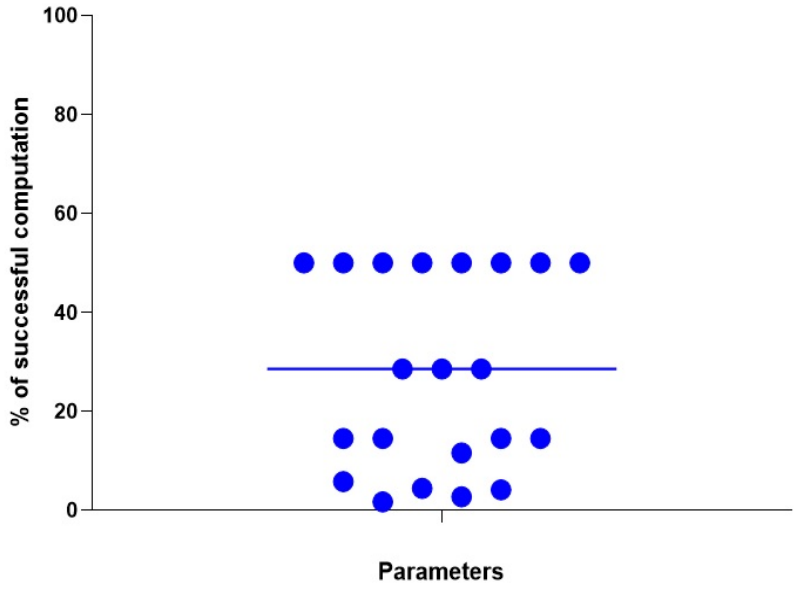


Figure 5. ***Probability of successful computation for the 21 parameters of the analyzed network with each parameter treated as independent.***

If a parameter has 3 or more settability paths, the probability of successful computation is equal or lower than 50%, respectively. However, the analyzed parameters are *dependent* parameters: their settability depends on having set other parameters to one value instead of another. As Figure 4 shows, in the Longobardi & Guardiano network, the first parameter that has 3 paths is P10. For 3 paths, the probability of successful computation on the first try is 50%. The second parameter that has 3 paths is P11. If this is taken as an independent event, the probability of successful computation is again 50%. However, if one wants to calculate the probability of a second successful computation under the assumption that the first parameter was computed successfully in one try, the conditional probability of successful computation in this second step is 25%. Table 3 shows that by the time the fifth parameter with 3 or more paths is encountered, the conditional probability of successful computation is 1.7%.

Table 3. ***Conditional probability of successful computation in one attempt (Longobardi & Guardiano network).***

Parameter	**Ways of computation without loop**	**Ways of computation with one loop**	**Conditional probability of successful computation**
P10 - 3 paths	6	6	50%

P11 - 3 paths	6	6	25%
P22 - 3 paths	6	6	12.5%
P23 - 3 paths	6	6	6.25%
P24 - 4 paths	24	60	1.7856%
P29 - 4 paths	24	60	0.5101%
P40 - 3 paths	6	6	0.2550%
P44 - 3 paths	6	6	0.1275%
P45 - 6 paths	720	4230	0.01858%
P46 - 6 paths	720	4230	0.00269%
P47 - 6 paths	720	4230	0.00039%
P48 - 6 paths	720	4230	0.000057%
P49 - 7 paths	5040	38262	0.00000662%
P50 - 3 paths	6	6	0.00000331%
P55 - 3 paths	6	6	0.00000165%
P56 - 15 paths	1.30767e+12	2.79027e+13	0.0000000740%
P57 - 4 paths	24	60	0.0000000211%
P58 - 16 paths	2.09228e+13	4.82395e+14	0.000000000878%
P60 - 23 paths	2.5852e+22	9.0699e+23	0.0000000000243%
P61 - 12 paths	479001600	7751595852	0.00000000000141%
P62 - 36 paths	3.71993e+41	2.13604e+43	0.00000000000002%

These findings raise concerns about computability, even if one assumes that the learner can somehow keep track of the fact that they have run into a loop, hence know that the computation should be re-run. This is highly pertinent in the context of a memoryless parsing process that knows only the current state. To explain the process from the learner's perspective, if on the first random selection of a path, A is chosen out of a set of paths ABCD, when the repetition of A occurs in the fourth selection (i.e., ABCA), the computation runs into a loop. Of course, the program that simulates the process keeps track of this possibility and flags it as a loop, because it was designed to do so. Yet the learner, who is equipped with a memoryless parser that lacks this feature, has no way of remembering which option was selected in the first/n^{th} random selection of a path. If the learner keeps track of the current state, ABCC can be recognized as a loop and be avoided, but ABCA cannot. In other words, the parser is oblivious to the fact that it runs into a loop more often than not.

Even if we endow the parser with the ability to recognize a loop and rerun the computation, concerns about computability are not sidestepped. The reason boils down to how the numbers of successful and unsuccessful computations were calculated above. It is important to stress that the developed program treats the presence of a single loop as an instance of unsuccessful computation. This means that if a parameter is settable on 4 paths ABCD, the order ABCA is a possible outcome that the program counts as unsuccessful, but ABCAB or ABAB are not possible outcomes for the program. Put another way, the program is purposely designed to count the event of falling into one single loop as the only case of unsuccessful computation, but in reality, the number of computations that involve a loop are infinite.

Restricting the possible number of unsuccessful computations by limiting the number of the random selection of paths to the number of paths (i.e., if a parameter has 4 paths ABCD, the learner is allowed to perform only 4 events of random path selections, enabling ABAB as an unsuccessful outcome, but not ABABA, thereby limiting the number of unsuccessful computations) does not alleviate concerns about computability. Under this limitation, a parameter with 36 settability paths has a total of $n \cdot (n-1)^{n-1} = 36 \cdot 35^{35} = 3.97e + 55$ ways of computation. Consider the computations that do not involve a repetition (Table 3; $3.71993e + 41$). There are $3.96903e + 55$ ways of unsuccessful computation, which translates to a $9.3 \times 10^{-13}\%$ probability of (the settability relations behind) this parameter being successfully computed in the first try. If a parameter has this probability of successful computation, the expected number of unsuccessful computations before a successful one occurs is:

$$E = \frac{1 - p}{p} = \frac{1 - 0.000000000000009372397825}{0.000000000000009372397825} = 1.06696e + 14$$

In other words, it is expected that more than 106 trillion unsuccessful computations will occur before a successful computation takes place.

To put the obtained results in comparison, we performed a second analysis using a different pool of data. Ceolin et al. (2021) present an expanded network that consists of 94 parameters from the nominal domain, covering 58 languages from 15 language families (Figure 6).

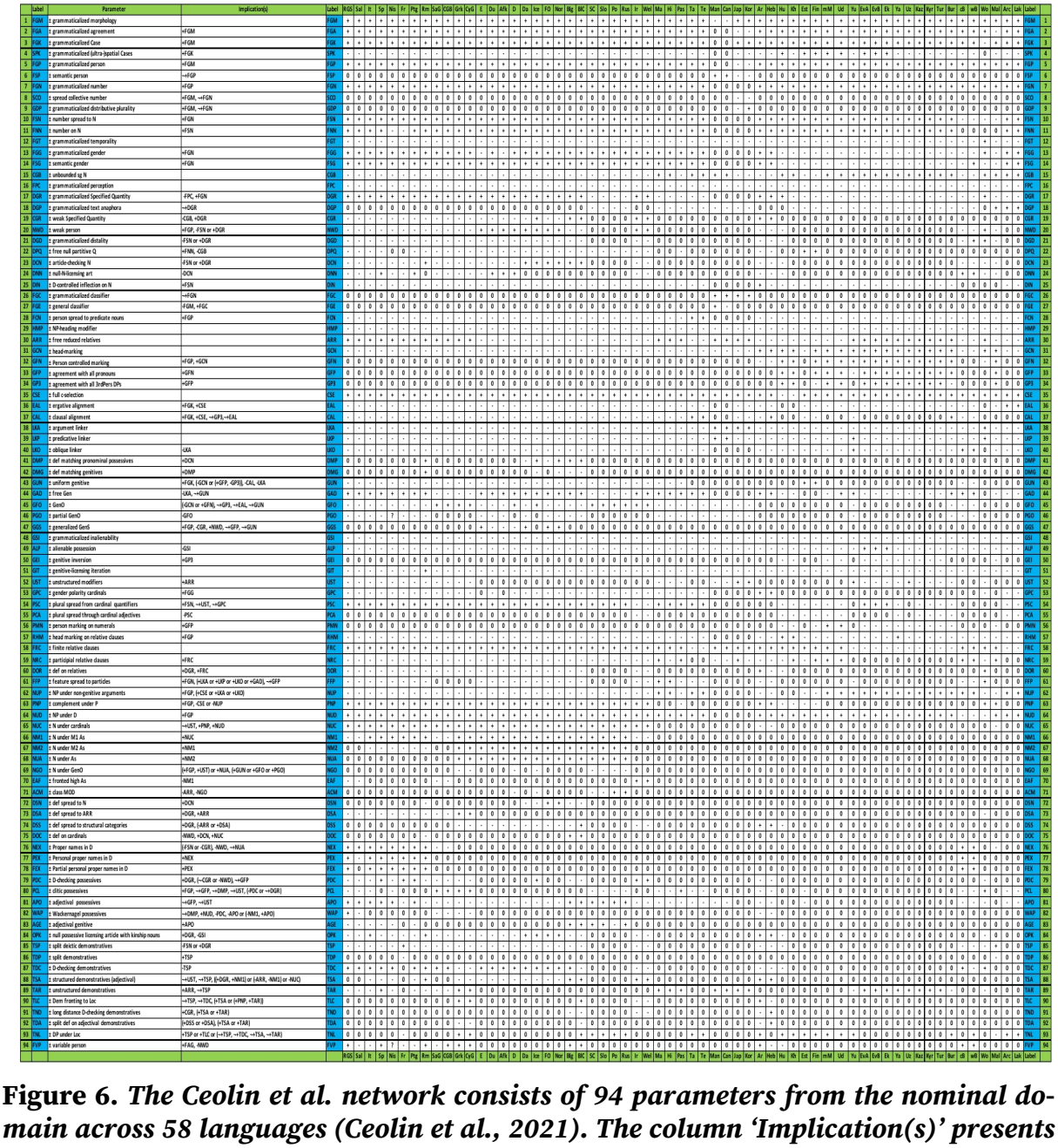


Figure 6. ***The Ceolin et al. network consists of 94 parameters from the nominal domain across 58 languages (Ceolin et al., 2021). The column 'Implication(s)' presents the settability path(s) on which each parameter is settable.***[1]

[1] For the purposes of our analysis, part of the logical expression behind the last parameter (i.e., FVP) was changed from FAG to FGA, following personal communication with Cristina Guardiano.

This network involves 82 dependent parameters, the settability of which depends on the setting of other parameters. Of these, 25 parameters are settable on 3 or more paths, and these are the ones we analyzed. Specifically, we converted the dependencies given in the 'Implication(s)' column (Figure 6) into mathematical expressions in the following way. If a dependency involves two parameters linked by ',' (i.e., ∧), both parameters must form part of every settability path behind this parameter, such that following Boolean logic this was expressed as a multiplication. If a dependency involves two parameters linked by 'OR' (i.e., ∨), each of the two parameters corresponds to a different way of reaching settability, so this was expressed as an addition. Example (4) illustrates the mathematical expression of a hypothetical example that has a structure that is found in the analyzed network (i.e., parameter 20, label: NWD, Figure 6).

(4) Parameter A = +B, +C or -D ⇔ 1 x (1+1) = 2 paths

In (4), the assumption is that parameters B, C, and D involve one settability path each. When this is not the case, the number of paths behind each parameter must be entered.

Table 4 presents the results of the mathematical expression of the relevant parameters in terms of settability paths as well as their probability of successful computation. For the latter, the Python program was used to perform the calculations.

Table 4. ***Dependent parameters with 3 or more settability paths and their (conditional) probability of successful computation (Ceolin et al. network).***

Parameter	Mathematical expression of the dependencies	Ways of computation without loop	Ways of computation with one loop	Prob. of successful computation	Conditional prob. of successful computation
P45 - 4 paths	(1 + 1) x 1 x 1 x 2	24	60	28.5%	28.5%
P46 - 4 paths	4	24	60	28.5%	8.16%
P47 - 4 paths	1 x 1 x 2 x 1 x 2	24	60	28.5%	2.33%
P61 - 5 paths	1 x (1 + 1 + 1 + 2) x 1	120	500	19.3%	0.45%
P62 - 3 paths	1 x (1 + 1 + 1)	6	6	50%	0.22%
P63 - 4 paths	1 x (1 + 3)	24	60	28.5%	0.064%
P65 - 4 paths	1 x 4 x 1	24	60	28.5%	0.018%
P66 - 4 paths	4	24	60	28.5%	0.0052%

P67 - 4 paths	4	24	60	28.5%	0.0015%
P68 - 4 paths	4	24	60	28.5%	0.0004%
P69 - 41 paths	(1 x 1) + 4 x (2 + 4 + 4)	3.34525e+49	2.2083E+51	1.4%	0.0000064%
P70 - 4 paths	4	24	60	28.5%	0.0000018%
P71 - 41 paths	1 x 41	3.34525e+49	2.2083e+51	1.4%	0.00000002%
P75 - 16 paths	2 x 2 x 4	2.09228e+13	4.82395e+14	4.1%	0.0000000011%
P76 - 16 paths	(1 + 1) x 2 x 4	2.09228e+13	4.82395e+14	4.1%	0.000000000047%
P77 - 16 paths	16	2.09228e+13	4.82395e+14	4.1%	0.0000000000019%
P78 - 16 paths	16	2.09228e+13	4.82395e+14	4.1%	0.00000000000008%
P79 - 3 paths	1 x (1 + 2) x 1	6	6	50%	0.00000000000004%
P80 - 8 paths	1 x 1 x 2 x 1 (3 + 1)	40320	375368	9.7%	0.0000000000000032%
P82 - 10 paths	2 x 1 x 3 x 1 + (4 x 1)	3628800	46253610	7.2%	0.000000000000000287%
P88 - 12 paths	1 x 1 x [(1 x 4) + (1 x 4) + 4]	479001600	7751595852	5.8%	0.0000000000000000167%
P90 - 13 paths	1 x 1 x (12 + (1 x 1))	6227020800	1.11471e+11	5.2%	0.00000000000000000088%
P91 - 13 paths	1 x (12 + 1)	6227020800	1.11471e+11	5.2%	0.00000000000000000004%
P92 - 39 paths	(2 + 1) x (12 + 1)	2.03979e+46	1.27643e+48	1.5%	0.00000000000000000000007%
P93 - 26 paths	1 + 13 + (1 x 1 x 12 x 1)	4.03291e+26	1.62279e+28	2.4%	0.00000000000000000000002%

As Table 4 suggests, two parameters in the analyzed network have 41 settability paths each. Repeating the analysis presented above for the Longobardi & Guardiano network (i.e., removing the one-loop restriction, but limiting the path-selection events to number of paths), a parameter with 41 settability paths has a total of $\mathrm{n} \cdot (\mathrm{n}-1)^{\mathrm{n}-1} = 4.95660e+65$ ways of computation. Subtracting the number of computations that do not involve a repetition (Table 4; $3.34525e+49$), there are $4.95659e+65$ ways of unsuccessful computation, which translates to a 6.7×10^{-15}% probability of (the settability relations behind) this parameter being successfully computed in the first try. Thus, the expected number of unsuccessful computations before a successful one occurs is:

$$E = \frac{1 - p}{p} = \frac{1 - 0.0000000000000067491}{0.0000000000000067491} = 14816817458844600$$

Succinctly put, it is expected that more than 14 quadrillion unsuccessful computations will occur before a successful one takes place.

Discussion

We have presented a previously unanalyzed aspect of the P&P approach that seems to entail an unrealistically cumbersome computational burden. We stress here that our report does not in principle repudiate the basic notion of parameters as emergent points of variation that build on innate principles, but rather the more specific conjecture that the infant is presented with an extensive predefined list of such parameters.

Parameters were proposed as a cognitive primitive that help organize and constrain the hypothesis space of a child trying to acquire language in an efficient way (Pearl & Lidz, 2013). Although the notion of parametric variation is theoretically well-formed and useful as a concept, previous research on the computation of parametric models of language acquisition has revealed various computability issues. For instance, it was found that the child would need to set about 30 parameters per second, throughout childhood, to assimilate a parametric model, with obvious consequences about computability (Levelt, 1974; Fitch & Friederici, 2012).

Other work on grammar learning revealed the *local maxima problem*: a learner may posit incorrect hypotheses about the target grammar G_t, forming a grammar G_s from which she can never move out, similar to an absorbing state in the theory of Markov chains (Gibson & Wexler, 1994). Related to this, the *learnability problem* refers to the fact that even if a path from G_s to G_t exists and there are salient cues that guide the learner towards the target, there is a high probability that the learner does not take this path, resulting in non-learnability (Niyogi & Berwick, 1996).

The *problem of low probability of unambiguous input* does not, strictly speaking, raise learnability concerns, but it does raise computability issues. According to this problem, given the scarcity of unambiguous input (i.e., there is no one-to-one correspondence between the surface properties of the input and the correct parameter values that generate G_t), the learning algorithm must wait for a sentence that is fully unambiguous before forming any G_s, yet these sentences have a very low probability of occurring (Sakas, 2000). Further, the notion of an unambiguous linguistic input also presupposes a robust and complex metacognitive, inferential state for the infant.

All these problems raise concerns that relate to forming hypotheses about a G_t in the process of parameter-setting, and not to determining settability. This means that they are problems that pertain not to the parametric model itself, but to the interaction between the input and the learner, and as such, they can be ameliorated under the

right conditions. For example, the local maxima problem can be solved if the learner can change more than one parameter setting when encountering input that is not predicted by G_s (Niyogi & Berwick, 1996). Similarly, the problem of low probability of unambiguous input has been sidestepped by suggesting that some sentences in the input function as signatures or unambiguous triggers; that is, they are analyzable only if the learner has selected the correct value for a parameter (Fodor, 1998; Yang, 2002). Focusing on setting relations, the conclusion is that under certain assumptions, parameter-setting *is* computable (Sakas et al., 2017). However, this state of affairs does not take into account the computability of settability relations.

Unlike problems of setting, problems of settability are *intrinsic* to the parametric model. To give a concrete example, the fact that one of the parameters analyzed in the previous section was found to have $3.96903e + 55$ ways of unsuccessful computation, even when restricting the possible number of loops to not exceed the number of possible path-selection events, is not a problem that the learner can overcome by using some particular learning strategy instead of another. No matter the strategy, the fact will remain that before one finds a way of checking these 36 settability paths without running into a loop, trillions of unsuccessful computations are expected to take place. Even under the unrealistic assumption that the child devotes only one second to each computation, execution would take 29,637,856,071 hours, or over 3 million years. This corresponds to the task of computing the settability relations behind a *single* parameter. It seems highly implausible that this amount of computation is entered into the task carried out by the child when acquiring language. To put the number in perspective, the discovery of the Ledi jaw that was recently added to the fossil record of the genus Homo places the earliest occurrence of recognizable Homo to 2.8 mya (Villmoare et al., 2015).

It may, of course, be possible for a deep learning approach to settability to reduce our large estimate of unsuccessful computations, in combination with external learning heuristics (of the kind we will discuss below). However, to our knowledge no such approach has been forthcoming in the literature, and in any event, it would likely necessitate a number of complex priors that may simply re-migrate settability difficulties to postulated AI algorithms that may have no cognitively plausible, implementational correlate (Marcus & Davis, 2021). The burden of proof in this respect lies with deep learning (and related) approaches, and we therefore leave this possibility to future research, in particular given that our approach here has been explicitly to model the computability of settability paths.

The results presented in the previous section demonstrate various problems. First, the memoryless parser cannot keep track of all the loops. Even if we endow it with this ability, the number of unsuccessful computations that run into a loop is in the thousands, and this is the case for parameters that have just 6 settability paths. Restricting the number of loops does not make the task feasible either. Importantly, the parameters that were analyzed represent only one domain of grammar: the nominal domain. One can imagine how much larger the task would be if more parameters are brought into the picture. In addition, setting these non-nominal parameters would

also rely on a number of complex, higher-order semantic and conceptual networks, whose developmental trajectory remains relatively elusive (Murphy, 2017). Second, the results suggest that a parametric approach to Universal Grammar is not feasible. Crucially, the results do not provide any kind of evidence against Universal Grammar itself, which remains a robust and necessary concept in some frameworks of language acquisition. The identified problems arise when one suggests that the grammatical relations described as parameters exist in the form of *interlocked primitives* in Universal Grammar. This entails that the results are also not informative about the grammatical properties that are described in the analyzed network: The parameters in the Longobardi & Guardiano and Ceolin et al. networks are correct in the sense that they faithfully represent some differences in the grammars of various languages. Both networks, beyond descriptive and typological evidence, are strongly supported by their phylogenetically plausible conclusions. Our results are informative about the computability of a key characteristic of parametric models: settability. This characteristic is the cornerstone of almost all parametric models of language acquisition, because it provides the answer to the logical problem of language acquisition. Parameters are meant to be understood as a built-in shortcut that aids acquisition (Pearl & Lidz, 2013), but this only happens when they are conceived as *interlocked* parameters, meaning that the setting of one parameter carries implications about the settability of others. If parameters were to be understood as millions of unrelated points of variation, the variation space would not be organized in specific ways, hence would not be an aid in acquisition.

These results challenge another long-standing assumption of parametric models: the *instantaneous* nature of acquisition. Chomsky introduced this metaphor with the aim of talking about an idealized version of development, one that abstracts away from specific stages, on the assumption that these stages are largely uniform and have no impact on the acquired grammar (Chomsky, 1975). Some research since then has proposed that this idealization can be treated as a viable research avenue for the topic of language acquisition (Cinque, 1989; Rizzi, 2000). The problem arises when the 'instantaneous acquisition' metaphor presupposes a Universal Grammar that is rich enough to justify the concept of rapid setting of innate primitives. In other words, the 'instantaneous acquisition' narrative *relies* on the existence of a structurally rich Universal Grammar that involves detailed parametric networks like the one analyzed here. Even if acquisition was instantaneous in the sense that the value of a parameter would be determined automatically without any of the parsing reported in acquisition models, the settability relations behind the dependent parameters would still need to be computed in a stepwise fashion. Unless a learner can perform some trillions of computations in an instant, acquisition cannot be viewed as an instantaneous process.

It is also important to note that the obtained results are informative about any given parametric model that postulates interlocked parameters. One may think that the multiple paths to the settability of a parameter in the two analyzed networks are an artifact of these specific networks, such that the settability problem would vanish if another network was examined. There are two reasons to believe that the opposite is true. First, the grammatical relations behind the parameters in the two networks are

correct and their faithful representation of cross-linguistic differences has never been challenged. Second, the neat binary branching of Figure 1 *is* an artifact of presentation. More specifically, it is an artifact of choosing some 'big' macroparameters and a few languages, oversimplifying and ignoring many intermediate points of variation. For example, some languages have both partial polysynthesis and null subjects, which is a combination Figure 1 does not permit. This possibility cannot be captured without adding more parametric nodes in the hierarchy. Once these nodes are added, Figure 1 will resemble the two analyzed networks. Overall, the obtained results confirm Chomsky's early disclaimer about instantaneous acquisition. In his words, the 'instantaneous acquisition' model "is surely false in detail, but can very well be accepted as a reasonable first approximation" (Chomsky 1967: 441-442).

In relation to the computability concerns our analyses raise, a reviewer notes that the formalization of the cross-parametric implications currently adopted in the networks represented in Figures 2 and 6 is not assumed to reproduce or simulate any learning process, and it is not based on any consideration concerning the potential computational effort made by the learner in processing this type of information. Thus, the possibility cannot be excluded that a different formalization of the same implicational network might produce different outputs that could also affect the settability relations we used in our analyses. Although this is true, the parametric inventories we analyzed are firmly grounded on solid descriptive, typological, and phylogenetic evidence (Crisma et al. 2020, Ceolin et al. 2021). As such, determining their computability is important. Naturally, if in future work the implicational network is altered specifically in order to be made computable/learnable, the observed computability concerns will be circumvented. Based on current knowledge, however, the fact remains that two examples of our best parametric inventories raise specific computability concerns at their present state of development.

These concerns beg two important questions about the scope of our results. A reviewer asks what would go wrong if the learner ignores the implicational network and just tries to opportunistically set parameters whenever possible. Relatedly, is it possible that our results do not raise computability concerns for P&P in general, but for one particular instantiation of a P&P model that involves a predefined list of options in the initial state of development? The answer to the first question is that the implicational network provides innate shortcuts that aid acquisition. Asking whether the learner could ignore it would be tantamount to asking whether we can ignore any other innate aspect of our biological make-up. More importantly, however, the learner has no reason to ignore it, because this implicational network is the glue that keeps together the parametric space. If we remove the glue, the learner is left to navigate an extremely large variation space without any shortcuts. This also answers the second question. As mentioned already, our results do not speak about Universal Grammar or the principles of P&P, hence it would be wrong to conclude that we cast doubt on P&P as a whole. We examined a specific aspect of its parametric component. In this context, the answer to the second question is that if we remove the implicational network from the picture, the computability issues we raised may be indeed sidestepped. However, this does not entail that we are left with a parametric model

that is free from computability concerns. In the absence of implicational relations, the learner faces the task of navigating an extremely large space of variation. It has been suggested that this large space of variation "brings to light a fatal weakness of the microparametric approach" (Huang & Roberts 2016: 321): Even as few as a hundred independent parameters would raise serious concerns about the realization of only a very small fragment of the set of possible grammars during the entire human history (Huang & Roberts 2016). In a nutshell, removing the implicational network from the picture possibly alleviates the computability problems we raised, but makes the model vulnerable to other issues. Of course, it is entirely possible that parametric models that do not suffer from any type of computability issues are developed in the future. At present, the most promising candidates are those that refer to emergent parametric hierarchies (Huang & Roberts, 2016; Biberauer, 2019). Once these proposals are developed in sufficient technical detail and mapped to cross-linguistic data, future studies that assess their computability will be possible.

Having shown that the process of grammar development does not correspond to fixing values of innate parameters, the question of how the child sets its target grammar becomes again relevant. Merging insights from different acquisition models (Yang, 2002; Chistiansen et al. 2009; Boeckx & Leivada, 2014; Fasanella, 2014; Westergaard, 2014; Yang et al., 2017; Chomsky, 2019), Figure 7 presents a sequence of seven processes that explain how the child extrapolates rules of grammar from the input. The aim here is to provide a detailed, biologically plausible account for this task, while assuming as few Universal Grammar-/language-specific primitives as possible. Figure 7 lists the tasks that the efficient learner has to perform in order to arrive at a target grammar G_t.

We will briefly describe the principles of computation that aid the learner in each of these tasks, as well as their neurobiological basis, effectively presenting the process of acquiring a G_t without resorting to postulating parameters. Importantly, we illustrate this model not to outline its specific algorithmic architecture, which deviates from the central critique and motivation we adopt here. Instead, we provide a general outline of an architecture that could feasibly be instantiated in a number of ways.

One crucial factor that unlocks the process of developing a G_t is very early prosodic information which helps eliminate logically possible (though unsubstantiated on the basis of the input) learning tracks. Therefore, the first step in the process of cracking the grammar 'code' is input segmentation, whereby the learner breaks a continuous acoustic or visuo-motor signal into a sequence of discrete, meaningless symbols that make up larger meaningful chunks. In order to go from continuous, unsegmented input to discrete elements, the learner must treat the input as meaningful across levels of linguistic analysis (Process 1 in Figure 7).

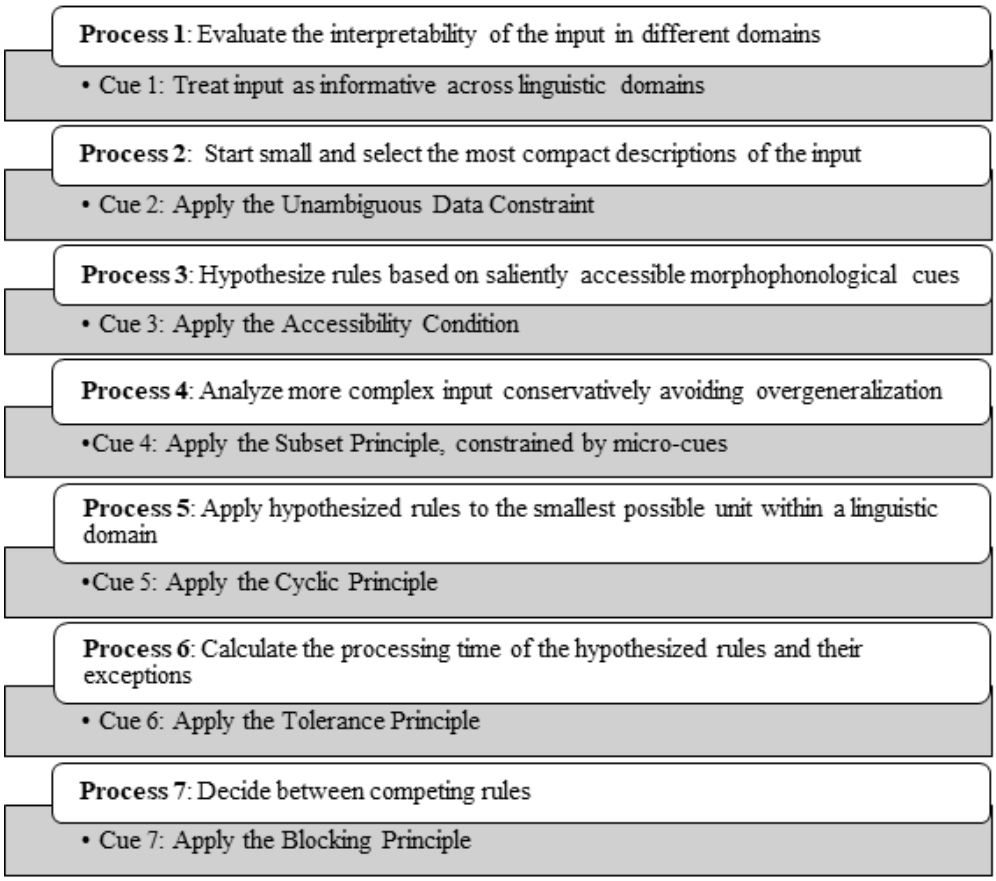


Figure 7. ***Processes and cognitive cues that are critical in developing a target grammar from the input.***

One crucial factor that unlocks the process of developing a G_t is very early prosodic information which helps eliminate logically possible (though unsubstantiated on the basis of the input) learning tracks. Therefore, the first step in the process of cracking the grammar 'code' is input segmentation, whereby the learner breaks a continuous acoustic or visuo-motor signal into a sequence of discrete, meaningless symbols that make up larger meaningful chunks. In order to go from continuous, unsegmented input to discrete elements, the learner must treat the input as meaningful across levels of linguistic analysis (Process 1 in Figure 7). More concretely, the computation progresses from forming statistical observations over phoneme distribution to deciphering word edges, segmenting morphemes, and then determining lexical categories (Christiansen et al., 2009). For spoken languages, the key to this process is the entrainment of the auditory cortex to different aspects of handling the acoustic signal, such as parsing at the syllabic level and integrating various cues while filtering background noise (Ding & Simon, 2014; Benítez-Burraco & Murphy, 2019; Murphy, 2015, 2020). For sign languages, cortical entrainment to the sign envelope is strongest at

occipital and parietal regions (Brookshire et al., 2017). After such initial entrainment, endogenous neural activity appears to "take over" and generate inferences about abstract structure, which we assume is the point at which grammatically relevant hypotheses can be made. This modality-independent stimulus-brain coherence underlies the extraction of probabilistic information from the input. Crucially, these processes presuppose a capacity to generate specific lexical categories but also a capacity to represent particular syntactic features that enter into structure-building operations; representations that seem unlike any other symbolic units in the primate world. In carrying out this process, the learner is initially guided by the Unambiguous Data Constraint, which leads them to select and focus on the simplest and cleanest possible data, mainly unambiguous matrix clauses (i.e., Process 2 in Figure 7; Lightfoot, 1991, 2020; Fodor, 1998; Pearl & Weinberg, 2007). This constraint can be viewed as the outcome of two hallmark tendencies of neural organization: the tendency to chunk long sequences and the tendency to organize/compress input in simple ways (Fonollosa et al., 2015; Christiansen & Chater, 2016; Chater & Loewenstein, 2016; Al Roumi et al., 2021). These tendencies are ubiquitous, but differentially manifested in accordance with the individual characteristics of spoken and signed phonology (e.g., single-segment words are rare in spoken languages, but common in sign languages, due to the different chunking strategies involved; Brentari, 1998; Emmorey, 2016). Having selected the relevant input, the learner then analyzes it by hypothesizing rules, based on saliently accessible morphophonological cues (Process 3; Boeckx & Leivada, 2014; Fasanella, 2014). According to the Accessibility Condition, grammatical properties of the G_t are determined by directly inspecting phonological and morphological properties of utterances (Fasanella, 2014). The speaker/signer analyzes an input chunk through hypothesizing a grammar G_i with a probability p_i. Depending on whether G_i matches the input from G_t, G_i is punished or rewarded by decreasing and increasing p_i accordingly (Yang, 2002).

Progressively, the learner tackles more complex input, but does so by avoiding overgeneralizations (Process 4). The Subset Principle guides the learner to generalize as conservatively as possible (Yang et al., 2017). Concerns that have been raised about the computational complexity of the Subset Principle (see Yang, 2016) can be sidestepped through the postulation of emergent (i.e., not innate) micro-cues. As minimal points of syntactic representation, micro-cues anchor the formed hypotheses in narrow domains of application, always on the basis of positive evidence (Westergaard, 2014). This anchoring renders wholesale, computationally costly comparisons of G_i and G_t unnecessary; a notion in line with recent developments in derivational syntactic theory (Chomsky, 2019; Murphy & Shim, 2020). Indeed, one of the implications of our results is that the initial hypothesizing on the part of the child of a large number of conflicting grammars is purely a stipulation from traditional psycholinguistic models, with no grounding in computability concerns. In a similar way that models of syntax no longer typically assume that multiple independent derivational representations of a specific tree are compared during sentence construction (as in early minimalist syntax), so too should language acquisition researchers push computational feasibility (and not competition between G_i and G_t) as a primary constraint on modelling.

Certain generalization tendencies do come into play (e.g., the Input Generalization, a computational bias that suggests that there is a preference for a property of a syntactic head to generalize to other heads, thus giving rise to harmonic patterns; Huang & Roberts, 2016), but they boil down to soft biases that do not translate into extensive overgeneralizations in child language. Their status as soft biases is also evidenced by the fact that they do not translate to absolute typological universals: Phylogenetic modelling has demonstrated that these generalizations are not uniform across language families (Dunn et al., 2011). Research into recently emerged sign languages corroborates this conclusion. There is some evidence for harmonic headedness patterns in the repertoire of first-generation signers of Al-Sayyid Bedouin Sign Language, but variation exists and the preference for one syntactic order over others becomes more stable progressively over different generations of signers (Sander et al., 2005).

Once the learner has hypothesized rules, a cognitive principle that minimizes the domain of application of these rules comes into the picture (Process 5). Similar to how the Subset Principle constrains generalizing across different morphosyntactic environments, the Cyclic Principle constrains the domain of application of the hypothesized rules. According to this principle, when one domain to which a rule can apply is contained in another, the rule applies first to the smaller domain and then proceeds to the wider one (Chomsky, 2019). From a biological perspective, this stepwise cyclical application of rules in grammar is concordant with the overall cyclical nature of auditory and visual perception, which has been linked to dynamic oscillatory activity in the brain (Ho et al., 2017). In addition, these notions seem amenable to ultimately being embedded within a framework of mature syntactic computation that calls upon demands of workspace construction; general resource restrictions on recursive, Markovian computations; limiting access to representational search; and related notions (Chomsky, 2019).

A key component of many acquisition models concerns the process that enables the learner to decide the productivity of a hypothesized rule in light of possible exceptions. The learner must perform some calculation that compares a list of candidates over which a rule applies and a list of exceptions to the rule (Process 6). The Tolerance Principle provides a calculus of the exceptions a learner can tolerate before abandoning a hypothesized rule as unproductive: Assume a rule R is productive over a set of items N only when the number of known exceptions *e* is smaller than the number of N divided by the natural log of N (Yang, 2002; Yang et al., 2017). The Tolerance Principle can also be shown to resolve the acquisition of English dative constructions, a perennial problem in acquisition research (Yang, 2017).

Last, the learner must be able to decide between different productive rules that may apply to the same item (Process 7). The Blocking Principle states that when two rules are available to realize a set of morphophonological values, the more specific one applies (Yang 2002). This ability to inactivate general rules in specific cases (e.g., not apply the regular rule for past tense formation in irregular verbs) provides the list of

exceptions that are necessary in the learner's effort to calculate the productivity of a hypothesized rule.

Overall, the list of processes in Figure 7 consists of some landmark cognitive principles that are operative in the process of language growth in the individual. Crucially, it shifts the focus of research to principles of computation, rather than triggered representational primitives. In addition, we have tried to emphasize the limitations on assuming models of idealized observers that choose either optimal or near-optimal hypotheses from an enormous list of explicitly entertained candidate settings. The model does not cover all aspects of acquisition; instead, it has an explicit focus on grammar, leaving other domains (e.g., the lexicon, pragmatics) unaddressed. Its scope is narrowed since our aim has explicitly been to account specifically for the process of cracking the grammar code without assuming innate parameters, in light of the computability problems presented above. Importantly, the program that performed the computations presented does not 'read' the linguistic properties behind the analyzed parameters; it only computes the various permutations between the settability paths behind them. As such, both the program that was used in the analysis of settability relations and the synthesis of cognitive principles that come into play in language acquisition can be embedded in wider contexts (e.g., by using the program to compute settability relations in other parametric models or by expanding the model in Figure 7 to include principles that are relevant in the process of lexical learning), eventually piecing together a more complete and biologically plausible account of the language acquisition process. At a minimum, our framework provides a (putatively) computationally tractable, and (seemingly) psychologically plausible scaffold around which implementational models can be built. We consider the account briefly outlined here to be ripe for future modelling research, in particular with respect to how the notion of computational tractability might map onto the development of general learning biases and computational principles of efficiency. Future research could expand on the list of parameters we have used and make more direct contact with models of cognitive and neural development (Crisma et al., 2020; Ceolin et al., 2020; 2021).

References

Al Roumi, F., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron, 109*(16), 2627–2639. doi: https://doi.org/10.1016/j.neuron.2021.06.009

Baker, M. (2003). Linguistic differences and language design. *Trends in Cognitive Sciences, 7*, 349–353. doi: 10.1016/s1364-6613(03)00157-8

Benítez-Burraco, A., & Murphy, E. (2019). Why brain oscillations are improving our understanding of language. *Frontiers in Behavioral Neuroscience, 13*, 190. doi: doi.org/10.3389/fnbeh.2019.00190

Biberauer, T. (2019). Factors 2 and 3: Towards a principled approach. *Catalan Journal of Linguistics, Special Issue,* 45–88. doi: doi.org/10.5565/rev/catjl.219

Boeckx, C., & Leivada, E. (2014). On the particulars of Universal Grammar: Implications for acquisition. *Language Sciences, 46*(B), 189–198. doi: doi.org/10.1016/j.langsci.2014.03.004

Boeckx, C., & Leivada, E. (2013). Entangled parametric hierarchies: Problems for an overspecified Universal Grammar. *PLoS ONE, 8*(9), e72357. doi: doi.org/10.1371/journal.pone.0072357

Brentari, D. (1998). *A prosodic model of sign language phonology*. Cambridge, MA: MIT Press.

Brookshire, G., Lu, J., Nusbaum, H. C., Goldin-Meadow, S., & Casasanto, D. (2017). Visual cortex entrains to sign language. PNAS, *114*(24), 6352–6357. doi: 10.1073/pnas.1620350114

Ceolin, A., Guardiano, C., Irimia, M.-A., & Longobardi, G. (2020). Formal syntax and deep history. *Frontiers in Psychology*, *11*, 488871. doi: https://doi.org/10.3389/fpsyg.2020.488871

Ceolin, A., Guardiano, C., Longobardi, G., Irimia, M. A., Bortolussi L., & Sgarro A. (2021). At the boundaries of syntactic prehistory. *Philosophical Transactions of the Royal Society B, 376*, 20200197. doi: https://doi.org/10.1098/rstb.2020.0197

Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization, 126*(B), 137–154. doi: https://doi.org/10.1016/j.jebo.2015.10.016

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1967). The formal nature of language. Appendix to E. Lenneberg's *Biological foundations of language*. New York: John Wiley and Sons.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Chomsky, N. (1981). Lectures on Government and Binding. Dordrecht: Foris.

Chomsky, N. (2019). Some puzzling foundational issues: The Reading program. *Catalan Journal of Linguistics, Special Issue,* 263–285. doi: doi.org/10.5565/rev/catjl.287

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39,* e62. doi:

doi.org/10.1017/S0140525X1500031X

Christiansen, M. H., Onnis, L., & Hockema, S. A. (2009). The secret is in the sound: From unsegmented speech to lexical categories. *Developmental Science, 12*(3), 388–395. doi: doi.org/10.1111/j.1467-7687.2009.00824.x

Cinque, G. (1989). Parameter setting in "instantaneous" and real-time acquisition. *Behavioral and Brain Sciences, 12,* 336.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. doi: https://doi.org/10.1017/S0140525X01003922

Crisma, P., Guardiano, C., & Longobardi, G. (2020). Syntactic parameters and language learnability. *Studi Saggi Linguistici, 58,* 99–130. doi: https://doi.org/10.4454/ssl.v58i2.265

Ding N., & Simon J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience, 8,* 311. doi: doi.org/10.3389/fnhum.2014.00311

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature, 473,* 79–82. doi: 10.1038/nature09923

Emmorey, K. (2016). Consequences of the Now-or-Never bottleneck for signed versus spoken languages. *Behavioral and Brain Sciences, 39,* e70. doi: https://doi.org/10.1017/S0140525X1500076X

Fasanella, A. (2014). *On how learning mechanisms shape natural languages*. Doctoral Dissertation, Universitat Autònoma de Barcelona.

Fitch, W. T., & Friederici, A. (2012). Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B, 367,* 1933–1955. doi: https://doi.org/10.1098/rstb.2012.0103

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry, 29,* 1–36. doi: 10.1162/002438998553644

Fodor, J. D. (2009). Syntax acquisition: an evaluation measure after all? In M. Piattelli-Palmarini, P. Salaburu, & J. Uriagereka (Eds.), *Of minds and language: A dialogue with Noam Chomsky in the Basque Country* (pp. 44–57). Oxford: Oxford University Press.

Fodor, J. D., & Sakas, W. G. (2005). The Subset Principle in syntax: costs of compliance. *Journal of Linguistics, 41,* 513–569. doi:

Fonollosa, J., Neftci, E., & Rabinovich, M. (2015). Learning of chunking sequences in cognition and behavior. *PLOS Computational Biology, 11*(11), e1004592. doi: https://doi.org/10.1371/journal.pcbi.1004592

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain*. Malden: Wiley-Blackwell.

Gibson, T., & Wexler, K. (1994). Triggers. *Linguistic Inquiry, 25*(3), 407–454.

Ho, H. T., Leung, J., Burr, D. C., Alais, D., & Morrone, M. C. (2017). Auditory sensitivity and decision criteria oscillate at different frequencies separately for the two ears. *Current Biology, 27*, 3643–3649. doi: https://doi.org/10.1016/j.cub.2017.10.017

Huang, C. T. J., & Roberts, I. (2016). Principles and parameters of Universal Grammar. In I. Roberts (Ed.), *The Oxford handbook of Universal Grammar* (pp. 306–354). Oxford: Oxford University Press.

Kazakov, D. L., Cordoni, G., Algahtani, E., Ceolin, A., Irimia, M-A., Kim, S-S., Michelioudakis, D., Radkevich, N., Guardiano, C., & Longobardi, G. Learning implicational models of universal grammar parameters. (2018). In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th International Conference (EVOLANG XII)*. Online at http://evolang.org/torun/proceedings/papertemplate.html?p=176.

Levelt, W. J. M. (1974). *Formal grammars in linguistics and psycholinguistics*. The Hague: Mouton.

Lightfoot, D. (1991). How to set parameters. Cambridge, MA: MIT Press.

Lightfoot, D. (2020). *Born to parse: How children select their language*. Cambridge, MA: MIT Press.

Longobardi, G., & Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua, 119*, 1679–1706. doi: https://doi.org/10.1016/j.lingua.2008.09.012

Manzini, M. R. (2019). Parameters and the design of the Language Faculty. Northern Italian partial null subjects. *Evolutionary Linguistic Theory, 1*(1), 24–56. doi: https://doi.org/10.1075/elt.00003.man

Marcus, G., & Davis, E. (2021). Insights for AI from the human mind. *Communications of the ACM, 64*(1), 38–41. doi: 10.1145/3392663

Murphy, E. (2015). The brain dynamics of linguistic computation. *Frontiers in Psychology, 6*, 1515. doi: 10.3389/fpsyg.2015.01515

Murphy, E. (2017). Acquiring the impossible: developmental stages of copredication. *Frontiers in Psychology, 8,* 1072. doi: 10.3389/fpsyg.2017.01072

Murphy, E. (2020). *The oscillatory nature of language*. Cambridge: Cambridge University Press.

Murphy, E., & Shim, J.-Y. (2020). Copy invisibility and (non-)categorial labeling. *Linguistic Research, 37*(2), 187–215. doi: 10.17250/khisli.37.2.202006.002

Niyogi, P., & Berwick, R. C. (1996). A language learning model for finite parameter spaces. *Cognition, 61*(1–2), 161–193. doi: 10.1016/s0010-0277(96)00718-4

Niyogi, P., & Berwick, R. C. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy, 20*, 697–719. doi: https://doi.org/10.1023/A:1005319718167

Page, K. M. (2004). Language learning: how much evidence does a child need in order to learn to speak grammatically? *Bulletin of Mathematical Biology, 66*, 651–662. doi: 10.1016/j.bulm.2003.09.007

Pearl, L., & Lidz, J. (2013). Parameters in language acquisition. In C. Boeckx & K. K. Grohmann (Eds.), *The Cambridge handbook of biolinguistics* (pp. 129–159). Cambridge: Cambridge University Press.

Pearl, L., & Weinberg, A. (2007). Input filtering in syntactic acquisition: answers from language change modeling. *Language Learning and Development, 3*(1), 43–72. doi: 10.1080/15475440709337000

Rizzi, L. (2000). *Comparative syntax and language acquisition*. London: Routledge.

Sakas, G. W. (2000). Modeling the effect of cross-language ambiguity on human syntax acquisition. *Proceedings of the fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, 61–66.

Sakas, G. W., Yang, C., & Berwick, R. C. 2017. Parameter setting is feasible. *Linguistic Analysis, 41*, 391–408.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic study in a new language. *PNAS, 102*, 2661–2665. doi: https://doi.org/10.1073/pnas.0405448102

Villmoare, B. et al. (2015). Early Homo at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia. *Science, 347*, 1352–1355. doi: 10.1126/science.aaa1343

Westergaard, M. (2014). Linguistic variation and micro-cues in first language acquisition. *Linguistic Variation*, *14*(1), 26–45. doi: https://doi.org/10.1075/lv.14.1.02wes

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, C. (2016). *The price of linguistic productivity. How children learn to break the rules of language*. Cambridge, MA: MIT Press.

Yang, C. (2017). Rage against the machine: evaluation metrics in the 21st century. *Language Acquisition: A Journal of Developmental Linguistics, 24*(2), 100–125. doi: doi.org/10.1080/10489223.2016.1274318

Yang, C., Crain, S., Berwick R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews, 81*, 103–119. doi: https://doi.org/10.1016/j.neubiorev.2016.12.023

Data, code and materials availability statement

The code is provided in the Appendix.

Authorship and Contributorship Statement

EL was involved in conceptualization of the research, data analysis, and data curation, and wrote the first draft of the manuscript. EM was involved in writing and editing the draft manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We thank the editor Brian MacWhinney and three anonymous reviewers for the useful feedback they provided. We are also grateful to Cristina Guardiano who answered questions about the analyzed pools of data, and to Cordian Riener for checking the mathematical aspects of the computation. This work received support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement n° 746652 and from the Spanish Ministry of Science, Innovation and Universities under the Ramón y Cajal grant agreement n° RYC2018-025456-I (to EL). The funders had no role in the writing of the study and in the decision to submit the article for publication.

Appendixes

```
import math
def computablePaths(paths):
   return math.factorial(paths)
def notComputablePaths(paths):
   if paths in [0, 1, 2]:
      return 0
   else:
      notCompPath = 0
      for l in range(2, paths + 1):
         temp = 1
         for t in range(0, l - 1):
            temp = temp * (paths - t)
         notCompPath = notCompPath + temp * (l - 2)
      return notCompPath

def calculateProbability(compPaths,notCompPaths, paths):
   totalPaths = compPaths + notCompPaths
   probability = float(compPaths / totalPaths)
   print(f"The probability of a successful computation is {probability * 100}%");

def main():
   finish = 1
   while(finish != 2):
```

```
    print("-" * 50);

    print("-" * 50 + "\n");

    paths = int(input("Number of paths: "))

    compPaths = computablePaths(paths)

    notCompPaths = notComputablePaths(paths)

    print(f"For {paths} paths, there are:\nWays of successful computation: {compPaths}\nWays of unsuccessful computation: {notCompPaths} \n")

    calculateProbability(compPaths, notCompPaths, paths);

    finish = int(input("\nDo you want to calculate another probability? \n1.Yes 2.No\n\n"))

    print("\n");


if __name__ == "__main__":

  main()
```

License

Predictors of children's conversational contingency

David Pagmar
Stockholm University, Sweden

Kirsten Abbot-Smith
University of Kent, UK

Danielle Matthews
University of Sheffield, UK

Abstract: When in conversation, a child may respond to an adult's turn in different ways: by saying something that acknowledges what was previously said, saying something that furthers the topic of the conversation, saying something off topic, or by not saying anything at all. Different types of responses like these have been investigated with typically developing preschoolers and older children with autism but we still understand relatively little about what predicts their use. With a longitudinal sample of 40 Swedish-speaking five-year-olds, we carried out three studies investigating which factors, internal and external to the child, were the best predictors of the above four different aspects of children's conversational behaviour. In Study 1, we investigated the predictive value of broadly concurrent linguistic and cognitive measures and found that receptive vocabulary was related to appropriate conversation responses. In Study 2, we investigated the predictive value of environmental factors and found that later preschool entry was positively related to contingent responses in this relatively socially advantaged sample. Finally, in Study 3, we investigated the predictive value of social and cognitive factors measured in early development and found no reliable relations. Together, these exploratory studies suggest that different aspects of children's conversational skills may depend on strong lexical comprehension and may be facilitated by the caregiving environment.

Keywords: conversational contingency; pragmatic development; first language acquisition

Corresponding author: David Pagmar, Department of Linguistics, Stockholm University, Universitetsvägen 10C, 11418, Stockholm, Sweden. Email: david.pagmar@ling.su.se

ORCID ID: https://orcid.org/ 0000-0002-6665-7502

Citation: Pagmar, D., Abbot-Smith, K., & Matthews, D. (2022) Predictors of children's conversational contingency. Language Development Research, 2(1), 139–179. https://doi.org/10.34842/2022-511

General Introduction

During first language acquisition, several fundamental elements must fall into place: a grammar, a lexicon, and control over a modality that can carry a linguistic signal. A language-acquiring child must also acquire the ability to use these fundamentals in social interaction. The ability to successfully use language for the purpose of social interaction and also take context into account when interpreting language is termed pragmatics. Pragmatic ability is closely linked to peer likability ratings (e.g. Place & Becker, 1991), child mental health (e.g. Helland, Lundervold, Heimann & Posserud, 2014) and poor pragmatic ability is associated with poor behavioural outcomes (e.g. Mackie and Law, 2010). Broad measures of child pragmatic ability are most frequently obtained via parental and teacher completed questionnaires (e.g. LUI, CCC2). Such questionnaires include items measuring child conversational ability, which is arguably the most frequent expression of pragmatic ability in daily life and for this reason child conversational ability is the key focus of the current paper.

Conversational abilities include engaging in turn-taking, offering relevant contributions to the conversation, and signalling interest in the contributions of others. The ability to maintain a back-and-forth conversation in this manner is essential for making and maintaining friendships (e.g. Hazen & Black, 1989) as well as collaborating on problem-solving activities both in school and in the workplace. For this reason it is important to understand which cognitive and socio-cognitive abilities, and which environmental factors, relate to individual differences in child conversational ability.

While norms differ across cultures, there are types of behaviour that are essential in conversational conduct, the most crucial component being the ability to provide a conversation response which is not 'tangential' in topic. A second important component is the ability to add new but relevant information so that the conversation can move forward. We follow Bloom, Rocissano, and Hood (1976: 528) in referring to the combination of these key conversational components as 'conversational contingency'; they state that *contingent speech* is defined as utterances that share the topic of the preceding utterance and add information to it (1976: 528). When a conversation partner provides a 'non-contingent' response, as in the example below from the current dataset, this can derail a conversation.

Experimenter: *You will eat a lot of ice cream! You mustn't forget your toothbrush.*
Participant: *I saw a horse on our way here.*

The definition of conversational contingency was adopted by later papers directly examining naturalistic conversations between children and adult conversation partners (Tager-Flusberg & Anderson, 1991; Hale & Tager-Flusberg, 2005a; Capps et al., 1998; Nadig, Lee, Singh, Bosshart & Ozonoff, 2010; Abbot-Smith, Matthews, Bannard, Nice, Malkin, Williams & Hobson, in prep) as well as by a study of semi-structured verbal interaction between typically-developing four- and five-year-olds and adults (Blain-Briere et al., 2014) and various studies of conversations between peers (e.g. Hazen & Black, 1989;

Kemple, Speranza & Hazen, 1992). Certain other studies have not utilised the term 'contingency' per se, but have examined the closely related phenomenon of 'connected' conversational responses - i.e. where the child's statement is logically related to the preceding statement and the back-and-forth conversation continues for a number of turns (e.g. Slomkowski & Dunn, 1996).

Past studies have put emphasis on different aspects of conversational behaviour, sometimes focussing on specific types of 'error' including going off topic (Hale & Tager-Flusberg, 2005b) or not responding at all (Capps, et al., 1998). Though both of these behaviours, going off-topic and not responding at all, can be considered less desirable conducts of a conversational partner, they do differ from each other. Non-contingent responses are potential contributions for someone else to follow up on, while a person that is not responding at all is basically opting out of the cooperative principle (Grice, 1975) all together. Also, these responses may be driven by very different cognitive factors. For example, not responding might logically be related to core language and the ability to formulate a response, in that a child must not only follow the conversational topic and realise what would be an appropriate contribution to the activity, but also have the means of producing a contribution and doing so in a timely fashion. It is possible that a child grasps the first two mentioned steps, but is having difficulties moving forward from there. In contrast, in order to produce a non-contingent response a child needs to have access to at least a certain level of vocabulary and morpho-syntax.

The aim of the current paper was to simultaneously look at these four related, but conceptually separated, conversational behaviours in children's responses to their interlocutor:

I. to add information and further the topic
II. to acknowledge what was previously said (whether it furthers the topic or not)
III. to respond without acknowledging the previous turn
IV. to not respond at all

We know that children will become increasingly sophisticated in their conversational strategies during the transition from preschool to school (Wanska & Bedrosian, 1985), but a pressing question remains unanswered: which factors allow children to develop the use of which conversational behaviours? By investigating I and II separately, we can see to what degree different correlates agree with the ability to specifically add new information to a conversation, and to what degree these correlates agree with the ability to acknowledge one's interlocutor in general.

Previous studies on conversational development have examined the role of formal language (e.g. vocabulary and/or grammar) and social cognition in typical and atypical development (Abbot-Smith, Matthews, Bannard, Nice, Malkin, Williams & Hobson, in prep; Abbot-Smith, Matthews, Malkin & Nice, 2021; Capps, Kehres, & Sigman, 1998; Hale & Tager-Flusberg, 2005; Bishop & Adams, 1989). Thus, Slomkowski & Dunn (1996) found that average length of preschool children's connected conversational turns in peer interaction, as well as the average length of play episodes and pretend episodes, were

positively related to performance on tasks of perspective-taking and false-belief (see also Bernard & Deleau, 2007:453, who did not examine observed conversation, but conversational perspective-taking). Likewise, Blain-Brière, Bouchard, & Bigras (2014) investigated the role of executive functions (self-control, inhibition, flexibility, working memory and planning) and observed that higher inhibition skills were correlated with a decrease in talkativeness and assertiveness, and that children with a high working memory capacity were more likely to formulate contingent answers (for further review of research on the relationship between pragmatic development and individual differences in language, social cognition and executive function, see Matthews, Biney, and Abbot-Smith, 2018).

Most studies, in contrast to those just mentioned, that address the connection between pragmatic development and other developmental factors, rarely assess direct measures of conversation. Another noteworthy exception is Hoff-Ginsberg (1998), who included both child internal (core language skill) and external factors (birth order, SES) when examining the development of conversation skill in younger children, aged 1;6–2;6. She found that first borns exhibited more advanced lexical and grammatical development, while later borns were more advanced in some types of (routine) conversational response. These results could indicate a division between conversational skill and core language development, or at least that they are not entirely dependent on each other. The children participating in this study were very young and studies on older children are needed to further examine these relationships with different types of conversational behaviour.

Other studies have explored the relation between the caregiving environment and the development of conversation in both typical and atypical development (e.g., Conti-Ramsden, Hutcheson, & Grove, 1995). Tomasello, Conti-Ramsden, & Ewert (1990) have suggested that the secondary caregiver (in their study, often the father) might prepare the child for communication with less familiar adults. A study of French toddlers similarly suggested a benefit of out-of-home daycare for some conversational behaviours (Marcos et al., 2004). Any relationship with the caregiving environment could of course be bidirectional. Indeed, in a study on three young children (1;9–2;6), Hoff-Ginsberg (1987) suggested that the conversation skill of the young child in turn affects the language learning environment.

Overall, while many studies suggest that different types of conversational behaviour are related to children's social and cognitive abilities as well as their caregiving environment, research in this area is still in its early stages. Thus, we conducted three studies, using data from one longitudinal data set, to explore the relationship between both child-internal and child-external factors and direct measures of four conversational behaviours. We examined two 'positive' behaviours: contingent responses (where the child adds to the conversation by contributing to the topic) and a broader category of appropriate response (where the child acknowledges the prior turn, but not necessarily with new information). We also looked at two types of 'error' that have received attention in the clinical literature: responding off-topic and not responding at all.

All studies were based on a preexisting Swedish longitudinal data set, the MINT project, with a conversational outcome measure at the age of 5;0 created by analysing

semi-naturalistic conversation. The measures of conversational behaviour were added to that dataset specifically for the current studies. The choice of predictor variables and sample size was constrained by the available dataset. While the studies are exploratory in nature, we nonetheless pre-registered all studies (osf.io/ah23m) and made hypotheses where theoretically appropriate.

We will present three pre-registered studies, each exploring how a set of predictor measures relate to each of the four types of conversational behaviour of interest. All analysed data stems from the same aforementioned data set. Study 1 was concerned with broadly *concurrent* measures of the child's ability to act in the world (measures of core language, conduct problems, curiosity). Study 2 was concerned with *environmental* factors: SES, birth order and daycare. Finally, Study 3 investigated whether developmentally earlier core language, social cognition and/or memory *longitudinally* predicted each of the four types of conversational behaviour.

General Method

Preregistration

The variables, hypotheses, and planned analyses for all three studies were pre-registered on Open Science Framework (https://osf.io/ah23m) after data collection, but prior to any analysis. Analysis scripts can also be found on OSF.

Participants

The sample consists of 40 Swedish speaking children (19 girls). Each child was at the age of 5;0 at the time of the recording of the conversational data (observed within two week from their birthday). All participating children were part of the longitudinal study MINT (MAW2011.007). Higher education was overrepresented among the parents of participating children, with 78% percent having studied at University level. Observations were made within two weeks of the child turning any specific reported age. A child from the MINT study was included in the current study if: 1) there were available longitudinal observations of the child, 2) the child's first language was Swedish, and 3) there were no reports of atypical development. In the conversational data, the children contributed with a total of 3612 conversational turns.

Testing procedure

All children were participants in the aforementioned longitudinal study MINT. Therefore numerous developmental test results (presented in detail below, as well as in Tables 2, 4, and 6) and longitudinal data were available for each participating child. For the current study, semi-structured conversations between the 5-year-olds and a researcher (the first author) were recorded with three stationary cameras and one in-action camera, worn by

the researcher. The children had met and interacted with the researcher on several previous occasions. For each child, we selected 10 minutes of conversation from the conversation partner's initial statement. All conversations were recorded in the same interaction laboratory at Stockholm University (PICTURE 1).

PICTURE 1: Four still photos taken from the video recordings of a session, showing all camera angles.

The child entered the interaction laboratory and was asked to sit down on a chair at a table. The researcher sat down on the opposite side of the table facing the child. The researcher then said the first out of 11 predetermined utterances. The reason for using predetermined utterances was to control the theme of the conversation and to make sure that each child would be given similar input from the researcher. Free interaction took place between the predetermined utterances.

Predetermined utterances

Below is a list of the 11 predetermined utterances that each participating child was exposed to during their recording session, translated into English from Swedish:

1. ”[NAME], how old are you?”
2. ”You know, Mo, Na, and Li, they live here in our lab, but tomorrow they will no longer be here”.
3. ”Where do you think they are gonna go?”.
4. ”They are going on vacation! Can you guess where they are going?”.
5. ”They will sleep in different places. Mo will sleep in a tree, Na will sleep on a roof, Li will sleep in a house”.
6. ”Mo will be gone for four days, Li will be gone for a few days, Na will be gone for a week, that's seven days. Who do you think will come home first?”.
7. ”They packed their bags this morning. Do you have a bag?”.
8. ”Do you know what happened when they were packing? They had a quarrel”.
9. ”Na thought that Mo had the plane tickets, but Mo hadn't seen the tickets”.
10. ”Na and Mo were really upset. They didn't know that Li had taken the tickets”.
11. ”Thank you [NAME], for talking to me about our friends!”.

Coding contingency and appropriate conversational behaviour

The conversational data was coded by the first author in accordance with a coding scheme for conversational contingency, developed by Abbot-Smith, Matthews, Malkin and Nice (2021), for which the coding manual is available on OSF (osf.io/q7wa4). Every turn that

the child took in response to the researcher during the conversation, both following the predetermined utterances and under the free interaction, was categorised into four basic categories:

> *contingent*,
> defined as an appropriate, informative and on-topic response to the experimenter's statements and questions,
>
> *non-contingent*,
> defined as a utterances that do not maintain the topic of the experimenter's statements and questions,
>
> *minimal response,*
> defined as utterances with little semantic weight, such as "Yeah" or "Wow". One-word utterances are normally coded as minimal, also imitative responses repeating what was just said,
>
> *other*,
> defined as responses on the part of the child or the experimenter that do not fit into any of the other categories, including laughter, inaudible responses, not-easily categorised responses, topic shifts following minimal responses from the researcher,

In addition to the categories listed above, *Missing turns* were also coded. A missing turn was coded when (i) >2 seconds had passed after the experimenter's turn, (ii) the child was not offering any vocal or gestural response, and (iii) the experimenter once again took a turn.

The categorical definitions above share similarities to previous coding schemes of children's adjacent and contingent responses. In the original definition from Bloom et al., (1976) a contingent response was defined as being "a response which, first, shared the same topic as the preceding utterance and, second, added information to the preceding utterance". In contrast, Blain-Brière, et al., (2014) did not include requirements for the response to be informative to be categorised as contingent, but that the utterance should be an "adequately respond to a request by the interlocutor". In the current paper, we followed Abbot-Smith et al.'s coding procedure in emphasising the second part of the definition, which meant that single word utterances and other utterances that did not add information (e.g. *did you?*) were excluded from the category of contingent utterances. The original papers that used this concept (Bloom et al., 1976; Tager-Flusberg & Anderson, 1991) outlined distinct sub-types of contingent responses, which both elucidates distinct ways in which they may be considered relevant to the preceding response and also explains how a response may be relevant but may nonetheless simultaneously 'move the conversation on'. One subtype was termed 'expansion' by Bloom et al. and involved adding information and content. The second subtype was termed 'alternation' and involves adding information which opposed the truth value of the preceding utterance (e.g. Mother: *this is a man?*, Child: *no, it's a lady*). The third subtype was termed 'expatiation'

and is the type of utterance which both adds information to the topic and simultaneously introduces a new related topic (e.g. Mother: *oh I'm glad a black dog came along and saved the bunny*, Child: *no, hunter shoot him*). In the current study all of the subtypes would be categorised as contingent responses.

Aside from considering potentially 'optimal' contingent responses we were also interested to explore any kind of basically appropriate response. We considered *appropriate* any contingent response along with any 'minimal responses' (e.g. one-word responses, phrases such as 'Did you?'). This behavioural category thus covers all instances where the child acknowledged their conversational partner's turn - where the child signalled that they were listening and that they are part of the conversation.

In contrast to responding in an appropriate manner, some children quite frequently go off topic. This has been the subject of some considerable research in the literature on autism and we wanted to explore this behaviour in the current study also. Finally, some children simply do not respond at all on occasion and we considered predictors of this inability to generate a response.

Thus, the purpose of analysis, each turn was coded with respect to the following four binary outcome variables that capture conversational (in)appropriateness in four different ways:

Contingent turns: was the utterance contingent on the prior turn?
Appropriate turns: was the utterance a contingent or minimal response, i.e. acknowledged the experimenter's previous turn?
Non-Contingent turns: was the utterance non-contingent (going off-topic topic)?
Missing turns: was the prior utterance followed by no response at all?

The question of which factors would predict each of these categories of conversational behaviour are of course to some extent related. We chose to investigate each of them in their own right, since they allow us to conceptualise conversation in slightly different ways, and we can obtain potentially valuable information from each, especially since there exists no one universally agreed-upon measure of what makes for 'good' conversation. Thus, piecing the results from these four analyses together helps us obtain an idea of which cognitive factors are for which kinds of conversational behaviour. For example, working memory difficulties might be particularly likely to lead to non-contingent responses (because children simply forget the topic) whereas psycho-social difficulties might more likely to predict null responses and formal language ability might be more likely to predict contingent turns (since the child would be able to fluently generate them).

It is worth noting that minimal responses made up a large part of what the children produced during the conversations. These turns were often appropriate, especially as feedback signals. A contingent turn marks that a child is cooperative and is contributing something to the conversation, but a minimal turn also often marks cooperativeness. In the examples below, translated from Swedish, 1b, 2b, and 3b are all categorised as minimal responses.

(1a) Experimenter: *That would be so crazy!*
(1b) Participant: *I know!*

(2a) Experimenter: *Na will live on a roof…*
(2b) Participant: *A roof?*
(2c) Experimenter: *…and Mo will live in a house*

(3a) Experimenter: *They will not be here tomorrow*
(3b) Participant: *Hm, ok*
(3c) Experimenter: *Where do you think they're going?*

In 1b, the participant is smiling and nodding their head while making the utterance. In 2b, the participant raises their voice to mark surprise. In both 1b and 2b, the participants are marking that they are engaged in the conversation. It can, at times, be more appropriate to say something short rather than something long, and by repeating what someone else just said, you can signal that you were listening. In 3b, the participant does not add much to the conversation but there is a case for labelling the response "appropriate" when evaluating the participants' conversational behaviour. In contrast, consider the following example:

(4a) Experimenter: *…and my favourite is ice cream*
(4b) Participant: [missing turn]
(4c) Experimenter: *What's your favourite?*

In 4b, the participant's gaze is directed toward a stuffed animal and they do not signal any communicative act directed towards the experimenter. If we compare 3b and 4b, one of the examples is clearly more cooperative than the other. In 3b, there is a response and it is connected to the previous turn. It is important to note that the majority of minimal responses in our dataset are more resemblant of 1b and 2b, than of 3b.

Inter-rater reliability

Twelve and a half percent of the data (i.e. five children) were coded by another native speaker of Swedish, blind to how the data was coded by the first author. There was a very high degree of reliability (Cohen's k = .91). The high result is in line with previous contingency coding results, e.g. Hale and Tager-Flusberg (2005a) obtained an IRR of Cohen's kappa = .88 – 1.00 per transcript. Nadig et al. (2010) obtained IRR of Cohen's kappa = .92 for response type.

Data treatment and analyses

Descriptive statistics

The mean, standard deviation, minimum and maximum value was calculated for all measured variables, presented below. The data was examined for outliers, defined as observations beyond 1.5 interquartile range below the first quartile or above the third quartile. One outlier was found in the outcome measure *Non-contingent turns* (i.e, one child produced relatively very many of these responses compared to others) . This was not a case of measurement error and given the statistical models we employed we saw no reason for excluding it.

Correlational analyses

For each study, we first present a correlation matrix using Pearson's R to understand the simple relationships between each of the four measures of conversation and their predictors. For these analyses, each of the outcome variables was the sum of each measure of conversation for each participant.

Regression analyses

Four separate analyses were conducted, one for each investigated conversational response types. This was repeated for all three studies. We fitted multilevel logistic regression models using the lme4 package (Bates, et al, 2015) in R (R Core Team, 210614). We held each occurrence of a coded conversational turn in the data as the dependent variable (N = 3612), where a turn that corresponded to the outcome measure was ascribed the value of 1, and all other turns were ascribed the value of 0. These binary variables allowed us to ask: to what degree is the occurrence of specific response type (i.e. the specific outcome measure) dependent on the predictors, compared to any other type of turn in the data? For each of the three studies, we examined the influence of the study specific predictors over the separate conversational measures.

The model predicts the outcome of the binary dependent variable in terms of log odds (logits) as a linear function of the predictors (the fixed effects). For each model, we included random intercepts for participants. Each study included different fixed effects (outlined below) depending on the research question.

Transformations

All continuous predictor variables were transformed to z-scores for the statistical analyses. One binary predictor in study 2, Older sibling, was dummy coded.

Model build and predictor evaluation

For each study, multilevel logistic regression models were built. Each model predicted the binary outcome measurements (Contingent turns, Appropriate turns, Non-Contingent turns, and Missing turns) and included random intercepts for participants. We report marginal R^2 and conditional R^2 (Nakagawa et al., 2017) by obtaining all variance-components of the mixed models. Marginal R^2 is calculated by dividing the fixed effects variance by the total variance. Marginal R^2 indicates to what level the variance in the data can be explained by fixed effects only. Conditional R^2 is calculated by adding the random effects variance to the fixed effects variance and dividing the sum of both by the total variance. Conditional R^2 indicates to what level the variance in the data can be explained by the full model. Random effects for each model are also presented in APPENDIX A.

Model performance in regards to the conventional limit for disregarding effects, i.e. p-values, will be presented, as well as Odds ratios (Szumilas, 2010) for all predictors with 95% confidence intervals. An odds ratio (OR) of 1 represents neither outcome being more likely than the other as a function of the predictor. An OR >1 means increased odds as a function of the predictor, an OR <1 means decreased odds. The distance in decimals from 1 is to be interpreted as percentages, i.e. an OR of 1.25 means that the odds are increased by 25%, an OR of 0.75 means that the odds are decreased by 25% .

All predictors are evaluated through a likelihood ratio test using the anova function in R. The likelihood ratio test compares a model with n predictors to a model with less than n predictors, in terms of likelihood of the data. We exclude one predictor at a time from each model, and then compare the new model with the one including all predictors. The tests are conducted to evaluate predictor contribution and we report χ^2 and p-value from each test in table 3, 5, and 7. The AIC values from each run are presented in APPENDIX B.

Study 1: Preschool language ability, psycho-social wellbeing and curiosity

Study 1: Introduction

In our first study, we examined whether different aspects of children's conversational skills relate to three factors, the first being the child's vocabulary and grammar. Previous studies have found fairly consistent positive relationships between these measures and pragmatic abilities (see Matthews, et al., 2018, for a review, although note also Hoff-Ginsberg, 1998). The role of core language in conversational proficiency might be expected since a child with a large vocabulary who can easily control a variety of grammatical structures would be more likely to have the linguistic skill necessary to predict and plan turns in fluent conversation.

Second, we explored children's psycho-social well-being, which we expected may have a two-way relationship with the ability to engage well in conversation. A few studies have examined this somewhat indirectly (e.g. Helland, Lundervold, Heimann & Posserud,

2014; Mackie and Law, 2010)). Mackie and Law (2010) found that primary-school aged children who were clinically referred because they showed "behaviour that was causing concern at school" had significantly greater language difficulties than matched 'control group' children from the same schools. This between-groups difference was particularly marked for pragmatic language, which includes conversational ability. Similarly, Donno, Parker, Gilmour and Skuse (2010) found that the only language-related differences between children referred for behavioural difficulties and matched controls pertained to pragmatic and not to formal / core language. A large-scale study found that pragmatic language skill mediated the relationship between structural language, on the one hand, and behavioural difficulties, as assessed by the Strengths and Difficulties Questionnaire (SDQ) (Law, Rush, & McBean, 2014). However, none of these studies directly assessed conversational ability. We do so here, albeit with a non-clinical sample that did not contain a large number of children with behavioural difficulties.

Third, we explored the role of the children's curiosity. Epistemic curiosity is described as the desire to seek new information (Litman, 2008). We were particularly interested in epistemic curiosity in relation to conversational contingency because to respond contingently, one has to listen to and engage with what the conversation partner has just said. To achieve this, one needs to be open to new topics from external sources over and above one's own drive to talk about things pertaining to one's own habitual interests. Thus, we assumed that a child that is curious about their immediate surroundings, and generally seeks new information, might be more likely to engage in conversation and be interested in engaging with conversation topics which are set by an adult experimenter. In turn, we assume that a child that is more likely to engage in conversation will to a higher degree be exposed to, and have the opportunity to learn from, conversational norms, than would a child that is not as likely to engage in conversation.

In sum, in Study 1 we examined broadly concurrent relationships between our conversational measures on the one hand, and on the other hand formal language (as assessed by receptive vocabulary and morpho-syntax), psychosocial wellbeing (as assessed by the Strength and Difficulties - SDQ - questionnaire) and epistemic curiosity (as assessed by parent-report). We predicted that vocabulary, morpho-syntax and epistemic curiosity would be positive predictors of *Contingent* and *Appropriate turns*, and negative predictors of *Non-contingent* and *Missing turns*. We predicted that assessments of psychosocial difficulties would be a negative predictor of *Contingent* and *Appropriate turns*, and positive predictors of *Non-contingent* and *Missing turns*, and that all three would each explain unique variance.

Study 1: Method

Obtaining predictor measurements

All predictor measurements were obtained when the children were above the age of 3;0.

Vocabulary (PPVT)

The Peabody Picture Vocabulary Test, PPVT-4 (Dunn & Dunn, 2007) was conducted when the participants were at the age of 4;0. The test was adapted for Swedish participants (Ahlström & Ljungman, 2011). Because this measure has not been standardized on a Swedish sample, raw scores were used. We note that this measure was collected one year before the children's conversational data was collected. However, on the basis of Song, et al. (2015) we consider it likely that this measure would be fairly stable over this timeframe and we therefore choose to label the observed measure of receptive vocabulary at 4;0 as a broadly concurrent measure.

Grammar

Grammar was measured through an adapted version of a core language skill scoring scheme (Tonér & Gerholm, 2021), which takes into account (1) morphosyntactic accuracy score, calculated as % well-formed clauses and (2) syntactic complexity, defined as subordinate clauses per word token. The measurement was obtained from the study's conversational data. The predicates produced by a participant, following the first 10 of the experimenter's predetermined utterances, were analyzed and the number of inflections was counted.

Psycho-social wellbeing (SDQ)

The Strengths and Difficulties Questionnaire (SDQ) is a widely used tool for measuring children's mental health and psychopathology between the age of 4 and 16 (Goodman, 1997). It measures five subtypes of behaviors: conduct problems, emotional problems, hyperactivity, peer problems, and prosocial behaviors. The validity of an adapted version for children between the age of 3;0 and 4;0 has been examined with satisfactory results (Croft, et al., 2015). The participants' parents answered the SDQ questionnaire when the children were at the age of 3;6. The measurement included in the study is a composite of all five subtypes. For this measure, a higher score indicates greater psycho-social difficulties.

Epistemic Curiosity

This was measured with an adapted version of a parent-report questionnaire, answered by the children's caregivers (Piotrowski, et al., 2014:547). The participants' parents answered the questionnaire, translated from English into Swedish, when the children were at the age of 3;6. The measurement included in the study is a composite of reported answers.

Study 1: Results

Descriptive statistics

The mean and standard deviation, as well as the maximum and minimum observed values, of the four outcome measures are presented in Table 1.

Table 1. Descriptive statistics for the four conversational outcome measures, as well as for conversational turns labelled Other (i.e. turns that did not fall into any of the predetermined categories). Measures are presented with mean, standard deviation, maximum and minimum score.

	Mean	SD	Median	Min	Max
Contingent turns	21.7	10.8	19	5	50
Appropriate turns	51.8	17.8	51	21	87
Non-contingent turns	2.8	3.5	2	0	19
Missing turns	12.3	7.9	12	1	32
Other	27.1	13.3	28	6	59

The predictors for all models in Study 1 were receptive vocabulary, expressive grammar, psycho-social wellbeing, and curiosity. The descriptive statistics for the predictors in Study 1 are presented below in Table 2.

Table 2. Descriptive statistics for all predictors in Study 1.

	Mean	SD	Min	Max
PPVT	62.7	15.5	19	101
Grammar	16.9	2.6	12	24
SDQ	15	4	7	24
Curiosity	35.7	5.9	22	43

Correlational Analyses

Figure 1 below outlines which study 1 factors were correlated with each of the four conversational measures (*Contingent turns, Non-Contingent turns, Appropriate turns*, and *Missing turns*) and their predictors.

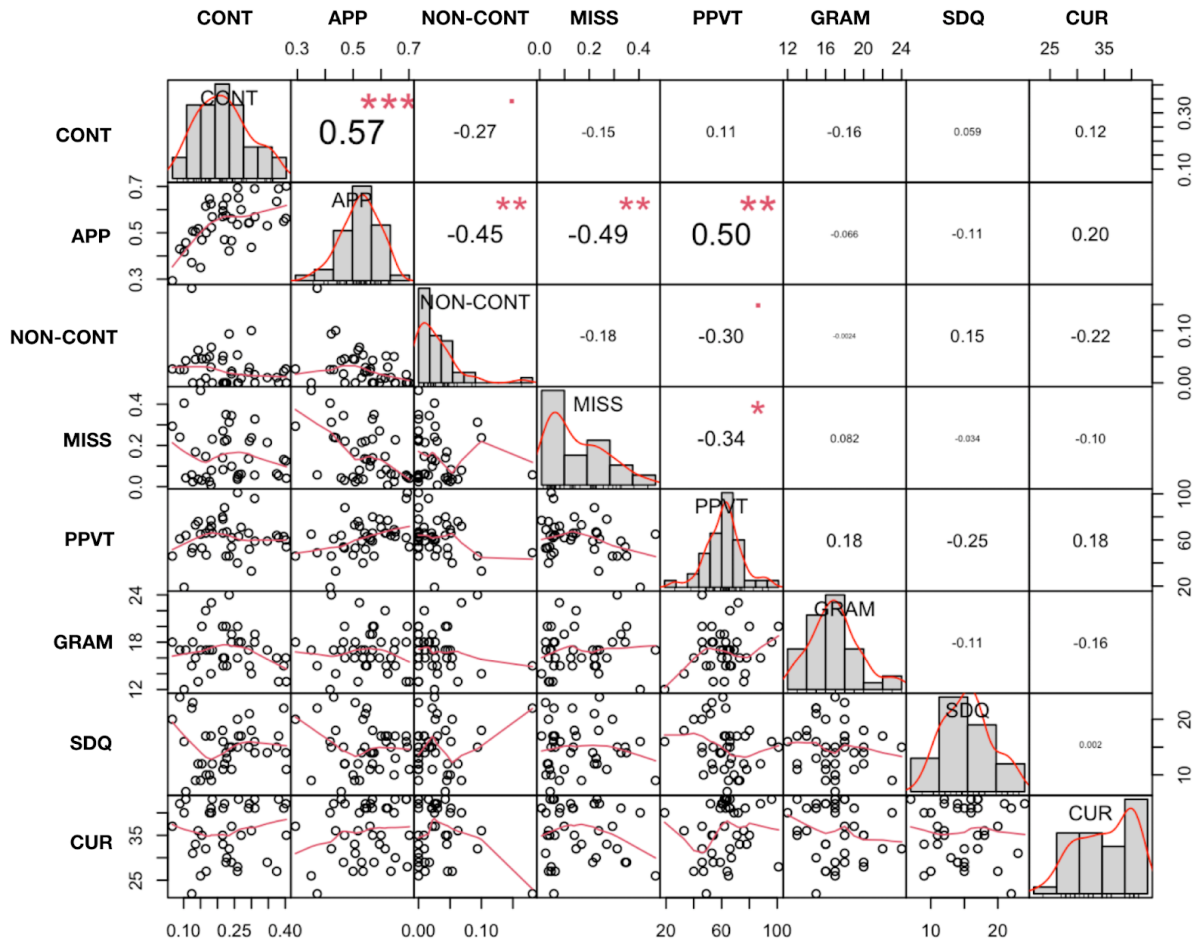


Figure 1. A correlation matrix showing pearson correlations between percentages of the four dependent variables per session: Contingent turns (CONT), Appropriate turns (APP), Non-Contingent turns (NON-CONT), Missing turns (MISS), and the predictors from Study 1 (standardized values): PPVT, grammar (GRAM), SDQ, curiosity (CUR).

Logistic regression analyses

Table 3 below reports findings for the fixed effects for each outcome variable in the logistic regression models (N = 3612), with χ^2 and p-values from the likelihood ratio test. Variance inflation factors were calculated and show no multicollinearity between predictors. Random effects for each model are presented in APPENDIX A.

Table 3. Fixed effects by dependent variable (Contingent turns, Appropriate turns, Non-Contingent turns, and Missing turns).

	CONTINGENT TURNS				APPROPRIATE TURNS				NON-CONTINGENT TURNS				MISSING TURNS			
	Est.	SE	χ2	p	Est.	SE	χ2	p	Est.	SE	χ2	p	Est.	SE	χ2	p
(Intercept)	-1.213	0.078	-	-	0.284	0.051	-	-	-3.89	0.217	-	-	-2.022	0.156	-	-
PPVT	0.090	0.084	1.138	0.286	0.221	0.055	13.337	<0.001***	-0.368	0.21	3.072	0.079 .	-0.368	0.167	4.522	<0.05 *
GRAMMAR	-0.078	0.081	0.902	0.342	-0.053	0.053	0.996	0.318	0.071	0.188	0.146	0.701	0.194	0.16	1.441	0.23
SDQ	0.029	0.08	0.133	0.714	-0.005	0.052	0.01	0.919	-0.012	0.19	0.004	0.945	-0.077	0.16	0.232	0.629
CURIOSITY	0.025	0.081	0.097	0.754	0.03	0.053	0.320	0.571	-0.03	0.2	0.023	0.879	0.036	0.161	0.05	0.822

Contingent and Appropriate turns

For Contingent responses, none of the predictors explained significant variance in the logistic regression model, all p:s > .29 (marginal R^2 = 0.003, conditional R^2 = 0.054) . As seen in Figure 2, the confidence intervals for the Odds Ratios for each predictor of contingent turns included 1. For appropriate responses, however, the vocabulary measure (PPVT) was a significant positive predictor (χ^2 =13.33, p<.001). While marginal R^2 for appropriate turns was 0.015 (conditional R^2 = 0.032), the Odds Ratios indicate that an increase in the PPVT vocabulary score by 15.5 points increases the odds for an appropriate turn by 24% [95%CI = 11%–39%].

Non-contingent and Missing turns

Vocabulary was a significant predictor of missing turns (χ^2 =4.52, p<.05) and showed a trend towards a negative relationship with *Non-Contingent turns* (χ^2 =3.07, p = 0.08). The relationship between vocabulary and the negative behaviour of missing turns mirrored the findings for appropriate turns; here an increase in vocabulary (PPVT) of 15.5 points decreases the odds of a missing turn by 31% [95%CI = 4%–51%]. No other measures reliably predicted negative conversation outcomes.

Odds ratios for Study 1

In Figure 2, below, we present the models in terms of Odds Ratios (Szumilas, 2010) with 95% confidence intervals .

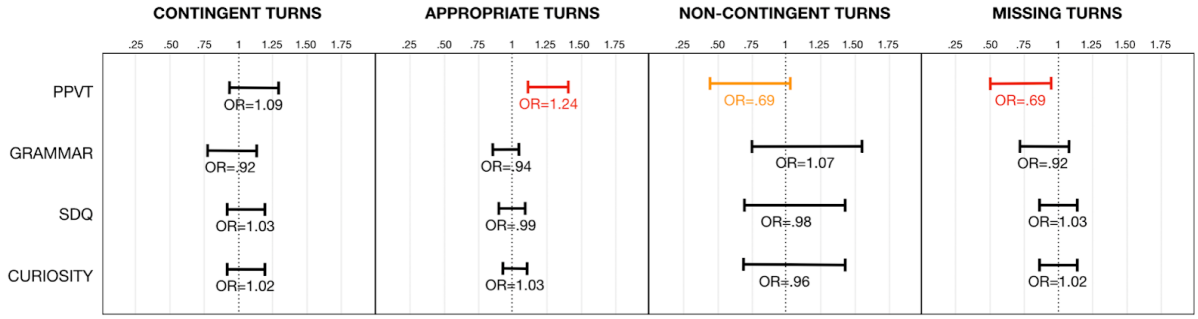


Figure 2. Odds ratios for the predictors in Study 1. The four different dependent variables are displayed in four columns (Contingent turns, Appropriate turns, Non-Contingent turns, and Missing turns). The predictors are displayed as rows. The odds ratios show how one unit in the predictor variable either increases or decreases the odds for the dependent variable to occur.

Study 1: Discussion

Vocabulary was a positive predictor for three of the measures of conversational ability. If a child had a relatively large vocabulary they were more likely to be able to generate a conversational response that was at least appropriate and they were less likely to simply not respond at all. Perhaps surprisingly, vocabulary was not a predictor of contingent responses. One might have expected that a strong vocabulary would be particularly valuable for generating contingent responses (as they tend to have more lexical content than minimal responses that do not move the conversation along). Contrary to our hypothesis, Grammar, SDQ, and Curiosity showed no significant relationships with any conversational measure.

Study 2: The language learning environment

Study 2: Introduction

Children's acquisition of formal language (vocabulary and morpho-syntax) is well-known to be influenced by environmental factors, particularly the quantity and quality of the language they hear directed to them (e.g. Hoff, 2003; Rowe, 2012). Factors such as Socio-Economic-Status (SES) or birth order are often used as proxies for the richness of child directed speech, however, to date very few studies have attempted to relate environmental factors to children's conversational ability. Study 2 thus included SES, time in day care and birth order in order to explore whether these environmental factors might predict the development of conversational proficiency.

Regarding SES, there is robust evidence for positive relationships between SES and vocabulary development in children (e.g. Huttenlocher, et al., 2010; Hoff, 2003; Rice &

Hoffman, 2015; Thornton et al, 2021). What is less clear is whether the positive direction of the relationship also holds for the development of conversational proficiency. On the one hand, parents from lower SES backgrounds have been shown to be less likely to follow in contingently on their own children's communications than do parents from middle-class backgrounds (e.g. McGillion, et al., 2017). This would suggest that children from lower SES backgrounds might have poorer conversational skills. Another aspect to consider is the environmental factors that may affect the conversational ability, like parental input. From observations of mother-child conversations, Hoff-Ginsberg (1991) found differences in the child-directed speech spoken in different settings between working class and upper-middle class mothers. Hoff-Ginsberg looked at several properties of maternal speech, e.g. number of utterances, utterances per minute, number of roots, MLU, % child utterances given topic-continuing replies, rate of conversation-eliciting utterances, rate of behaviour directives. When considering all settings, upper-class mothers scored higher in all categories, except for rate of behaviour directives. For specific settings, such as reading, all differences were not detectable. This also might suggest that children from lower SES backgrounds receive less exposure to conversational conduct compared to children from higher SES backgrounds.

On the other hand, there are suggestions from the work of Labov that higher SES children may not make better conversational partners, and indeed the reverse might even be the case in some respects (Labov, 1969). Hoff-Ginsberg (1998) found no reliable difference between mid- and high-SES when examining young children's conversational skills. In a recent study, Schulze and Saalbach (2021) looked at children's performance in a communication task and found no predictive value from parents' educational background or income. In the current study we explore the relation between conversational ability and SES operationalised as the mean income in the families' postcode areas.

Another aspect of the input which is often less considered is the language that children hear in different caregiving contexts, for example at home or at daycare. This might be particularly important in terms of learning how to hold a back-and-forth conversation. At preschool, children will be exposed to different language users including many peers and a range of caregiving adults. This might lead one to assume that an earlier start at preschool could result in better pragmatic ability. While the opportunity for peer-interactions has been explored to some degree in relation to how children learn to tell narratives (e.g. Küntay & Senay, 2003), to our knowledge there has been little exploration of this with relation to child conversational skills. One the one hand one might expect a similar advantage while on the other,, given the complexities of preschool quality and the tradeoff with alternative caregiving environments (see e.g. Burchinal, Roberts, Nabors & Bryant, 1996), there may also be reasons *not* to expect a simple positive relation between time spent in preschool and conversational skill. In one study of 27-month-old French children, Marcos et al (2004, p.145) found that there was a certain advantage for children who had daycare outside of the home or in terms of the amount of turns in conversation with their mother but not in terms of the thematic contingency of those turns on what their mother had said. On the other hand, NICHD Early Child Care Research Network (1999) looked at assessments of longitudinal mother–child interaction and found "small

but significant" results showing that more child-care hours negatively predicted two interactional components: child engagement and maternal sensitivity. If these findings persist beyond early childhood, the notion of less child engagement in interaction with parents might result in less interactional engagement overall. In the current study we explored whether starting nursery at an earlier age and spending more time there predicted better conversational contingency in Swedish 5-year-olds.

Finally, we were also interested in examining environmental effects driven by the presence of an additional sibling with whom the child has to share the parent's attention and language input. Previous findings show that first-born children are at an advantage in terms of expressive vocabulary size (Urm & Tulviste, 2016; Pine, 1995). As seen above, vocabulary is a positive predictor for conversational behavior, and therefore we might expect it to also be a positive predictor of conversational ability. However, while Hoff-Ginsberg (1998) also observed a first-born advantage for vocabulary, she simultaneously saw a trend in the opposite direction for conversational contingency, at least for 18- to 29-month-olds, which might be taken to suggest the two phenomena are somewhat separable. It appears these later-borns relied on what were coded as social *routines* to reply contingently to their caregivers more readily without taxing their more limited lexical resources. Such routine responses include saying things like "I don't know", "I can't", or "thank you" - responses that would be coded as a *minimal* turn, rather than a contingent turn in the current study (i.e., appropriate but not adding very much). When Hoff analysed the proportion of contingent responses that were expansions or expatiations (most similar to contingent replies in this study), first borns produced proportionally more such responses. The picture is thus somewhat mixed in toddlerhood. Nonetheless, when we consider development beyond toddlerhood, the literature on Theory of Mind development (e.g. Perner, Ruffman & Leekam, 1994; Hughes, 2011) might be taken to predict that having older siblings results in more advanced social cognition which could benefit conversation. In the current study we therefore also examined whether having an older sibling was associated with better conversational skill in 5-year-olds but we did not have a directional prediction.

In sum, in Study 2, we did not have a directional prediction for socioeconomic status (SES) or birth order. However we expected more time in day care to be a positive predictor for *Contingent* and *Appropriate turns,* and a negative predictor for *Non-contingent* and *Missing turns*.

Study 2: Method

The predictor measures for Study 2 were as follows.

Socioeconomic status

SES was measured in terms of the mean income in each families' postal code area because our participants' parents all came from very similar educational backgrounds.

Educational level was also recorded but the measure did not show enough variance, with a large majority of parents having undergraduate qualifications.

Preschool start

This measurement was assessed in terms of the child's age in weeks when they started attending daycare.

Preschool hours per week

The measurement consisted of the number of hours per week that the child attended day-care when aged 2;3.

Older siblings

Children who had one or more older siblings received a score of 1 and all other children received a score of 0.

Study 2 – Results

Descriptive statistics

Table 4 below presents the descriptive statistics for the Study 2 predictors that were continuous variables. 55% of our final sample had an older sibling.

Table 4. Descriptive statistics for predictors variables in Study 2: SES (represented by mean income in postal code area presented in Swedish crowns), Preschool start in weeks, and Preschool hours per week.

	Mean	**SD**	**Min**	**Max**
SES (income/postal code in SEK)	403793	87917	279199	648533
Preschool start in weeks	76	17	51	106
Preschool hours per week	34	7.6	7	46

Correlational Analyses

Figure 3 below outlines which Study 2 factors were correlated with each of the four conversational measures (*Contingent turns, Appropriate turns, Non-Contingent turns*, and *Missing turns*).

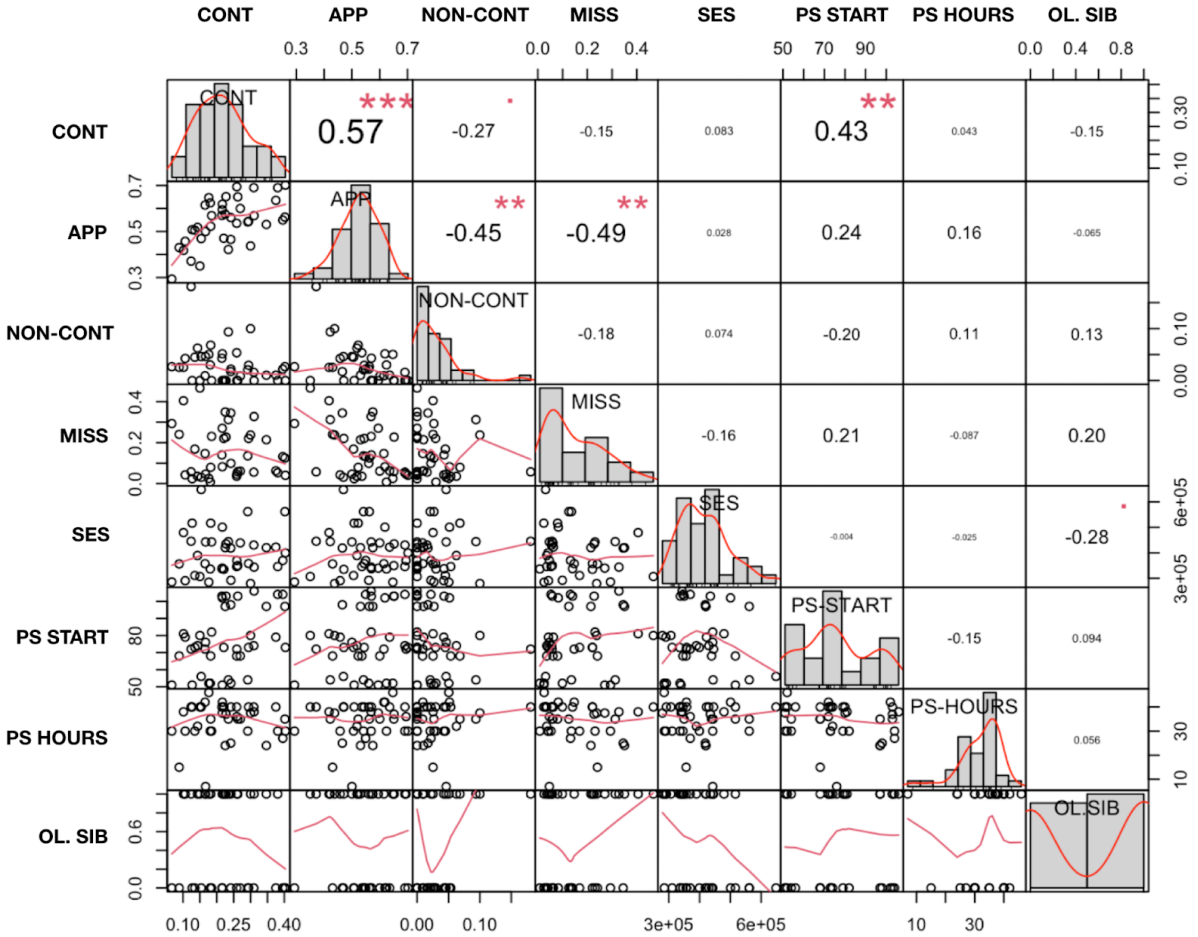


Figure 3. A correlation matrix showing Pearson's correlations between percentages of the four dependent variables per session: Contingent turns (CONT), Appropriate turns (APP), Non-Contingent turns (NON-CONT), Missing turns (MISS), and the predictors from Study 2 (standardized values): Socioeconomic status (SES), preschool start (PS START), preschool hours (PS HOURS), and older sibling (OL. SIB). For older sibling, we presented the point-biserial correlation coefficient.

Logistic regression analyses

Table 5 below reports findings for the fixed effects for each outcome variable in the logistic regression models (N = 3612), with χ^2 and p-values from the likelihood ratio test.

Variance inflation factors were calculated and show no multicollinearity between predictors. Random effects for each model are presented in APPENDIX A.

Table 5. Fixed effects in Study 2 by dependent variable (Contingent turns, Appropriate turns, Non-Contingent turns, and Missing turns).

	CONTINGENT TURNS				APPROPRIATE TURNS				NON-CONTINGENT TURNS				MISSING TURNS			
	Est.	SE	χ2	p	Est.	SE	χ2	p	Est.	SE	χ2	p	Est.	SE	χ2	p
(Intercept)	-1.130	0.104	-	-	0.303	0.089	-	-	-3.939	0.289	-	-	-2.173	0.232	-	-
SES	0.009	0.077	0.013	0.906	-0.029	0.065	0.208	0.648	0.088	0.195	0.202	0.653	-0.139	0.171	0.65	0.42
PS_START	0.224	0.071	8.671	<0.01 **	0.071	0.062	1.285	0.256	-0.312	0.204	2.37	0.123	0.243	0.159	2.2945	0.129
PS_HOURS	0.062	0.073	0.729	0.392	0.068	0.062	1.202	0.272	0.024	0.190	0.016	0.898	-0.093	0.160	0.337	0.561
OL.SIBLING	-0.177	0.15	1.347	0.245	-0.066	0.128	0.265	0.606	0.128	0.410	0.096	0.755	0.288	0.328	0.756	0.384

Contingent and Appropriate turns

Recall that contingent responses and appropriate responses were both positive measures of conversational ability. For contingent responses, there was a significant positive effect of preschool start in weeks (χ^2 =8.67, p<.01). While marginal R^2 for contingent responses was 0.015 (conditional R^2 = 0.053), the Odds Ratios (see Fig 4) indicate that starting preschool 17.6 weeks *later* increased the odds of a turn being Contingent by 25% [95%CI = 8%–44%].

Non-contingent and Missing turns

Recall that non-contingent and missing turns were both negative measures of conversational behaviour. For both non-contingent and missing turns, no predictors reliably explained variance in negative conversation outcomes (all p > .12).

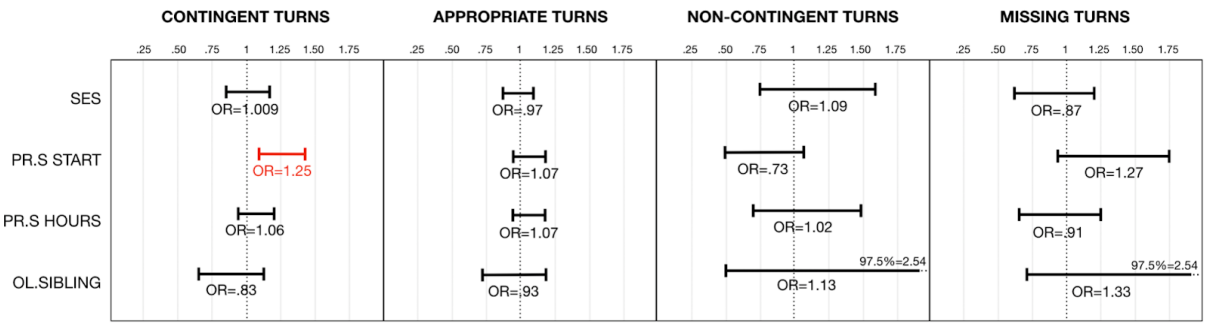


Figure 4. Odds ratios for every predictor in Study 2. The four different dependent variables are displayed in four columns (Contingent turns, Appropriate turns, Non-Contingent turns, and Missing turns). The predictors are displayed as rows. The odds ratios show how one unit in the predictor variable either increases or decreases the odds for the dependent variable to occur.

Odds ratios for Study 2

Below, we present the models in terms of odds ratios (Szumilas, 2010) with 95% confidence intervals (figure 4).

Study 2: Discussion

The results do not support the hypotheses that more time in daycare would have a positive effect on conversational contingency – if anything children with a *later* preschool start had an advantage in the number of contingent turns. Finally, neither SES nor presence of older siblings was associated with any conversation measure. In the case of SES we note that this measure did not have a high degree of variance. However, for sibling status, the sample was approximately evenly distributed regarding having an older sibling or not but this factor nonetheless showed no relationship with any of the four conversational behaviours. It might be that a finer grained analysis would reveal differences in *how* children are responding contingently (see Hoff-Ginsberg, 1998). We should also note that the current conversational measures are based on interaction with an adult not a peer, which might advantage first borns.

Study 3: Longitudinal examination of early vocabulary, short-term verbal memory, and imitation

Study 3: Introduction

To date, hardly any studies, to our knowledge, have explored whether children's conversational abilities can be longitudinally predicted on the basis of measures of their earlier cognitive and socio-cognitive development. In our third study, we explored whether children's appropriate conversational responding could be predicted on the basis of their earlier vocabulary, memory, and social cognition.

As we saw in study 1, children's conversational ability is associated with their broadly concurrent vocabulary. We do not know how stable this association is over time, however, and whether early vocabulary difficulties might be predictive of later conversational difficulties. Here we tested whether children's *expressive* vocabulary at age 2;3 was predictive of conversational ability when they were 5-years-old.

We also considered the role of short-term memory in relation to conversational ability. To provide contingent turns when taking part in back-and-forth conversation, besides keeping track of the conversation topic, one also needs to continuously keep in mind what an interlocutor just has said. The ability to maintain, manipulate and update information in short term memory is commonly referred to as working memory (Blakey, Visser, & Carroll, 2016) and has been found to correlate with conversational ability. Blain-Brière, Bouchard, & Bigras (2014) found that verbal working memory (Backwards Digit Span) related positively – with an effect size of 0.25 – to conversational contingency (and

was the only factor which correlated with contingency) in a sample of 70 typically-developing four- and five-year-olds. The memory variable we had available was a measure of phonological short-term memory (forward digit span) taken when children were 2;9. This measure did not involve manipulating information in memory (as the working memory measures noted above do) since this is difficult to assess at such an early age. Nonetheless, previous work has shown that phonological short-term memory capacity is an important predictor of vocabulary acquisition and word learning in both children (5-year-olds) and adults (Gathercole, et al, 1997) . We tested if early measurements of short-term verbal memory, taken at age 2;9, predict appropriate conversational behaviour.

Finally, to show appropriate conversational behaviour, it is arguably important to take the interlocutor's mental states into consideration. Such social cognition has often been measured by assessments of false belief which is arguably not necessary for many conversational interactions. We explored a more basic index of social cognitive ability: imitation. The imitation measure used in this study was part of the aforementioned pre-existing data set and was selected for inclusion as a marker of early social cognition, which we expected could pave the way for good conversational skills. Previous findings show that parental assessed measures of imitation show moderate explanatory value for variation in concurrent parental assessed conversation skill (Farrant et al., 2011), which prompts the question if such a relationship is detectable longitudinally as well. Meltzoff and Decety (2003) suggests that infant imitation provides the foundation for understanding that others are 'like me', i.e. have the same mental experience, and that it underlies the development of theory of mind and empathy for others. There is a large body of research showing links between action imitation and early communication development (e.g. Carpenter, Nagell & Tomasello, 1998; Carpenter, Tomasello & Striano, 2005; Zambrana, Ystrom, Schjølberg & Pons, 2013). These two abilities may be interrelated because a child that is inclined to imitate the actions of others, understands others' goals and means and is inclined to adopt them in purposive behaviour, of which conversation is an example. Findings from Nagy (2006) show that infants used previously imitated gestures to initiate communication, and although the study was concerned with very rudimental communicative actions, it exemplifies the notion of an agent observing an act, imitating the act, and later reproducing the act for their own communicative purposes. Previous studies have found that children with language impairments have greater difficulties than do well-matched neuro-typical peers with certain types of action imitation (Dohmen, Chiat & Roy, 2013). We therefore predicted that early imitation ability would predict later conversational ability.

We expected that early measures of children's vocabulary, memory and imitation would be positive predictors of conversational ability.

Study 3: Method

Vocabulary

Expressive vocabulary was assessed by the parental questionnaire SECDI-II, the normed Swedish translation of the McArthur-Bates Communicative Developmental Inventory (https://mb-cdi.stanford.edu/; Berglund & Eriksson, 2000, Larsson, 2014). This measure was chosen because direct measures of vocabulary are difficult to administer below the age of 3 years. The participants' parents answered the questionnaire every third month during the participants' first three years of life. We selected the measure which was obtained when the participants were at the age of 2;3 because the distribution showed variance without clear floor or ceiling effects.

Forward digit span

Participants were asked to repeat a series of random digits that the experimenter said, initially two at a time. The experimenter added one digit every other turn making the series of digits successively longer. The test was stopped after the participant made two errors in a row. The number of correctly repeated series of digits was counted. This measurement was obtained when the children were aged 2;9.

Imitation

Our imitation test is an adapted version of a longitudinal within-participants imitation task (Sakkalou, et al., 2013). In this task, the participant is prompted to imitate a test leader that is engaging in pretend play, making building blocks jump, building a tower with the building blocks, clapping hands, and putting the blocks into a bag. Each test part was scored as follows, no imitative action = 0, close to imitative action = 1, full imitative action = 2, for a potential maximum score of 8. The measurement was obtained when the participants were aged 1;0.

Study 3: Results

Descriptive statistics

The predictors for all models in Study 3 were early vocabulary, working memory (assessed via Forward Digit Span), and early imitation ability (assessed via the action imitation test). The descriptive statistics for the study 3 predictors are presented in Table 6 below.

Table 6. Descriptive statistics for all predictors in Study 3.

	Mean	SD	Min	Max
SECDI-II (at age 2;3)	319.2	152.6	24	653
Forward digit span	1.9	1.4	0	4
Imitation	2.8	1.4	0	6

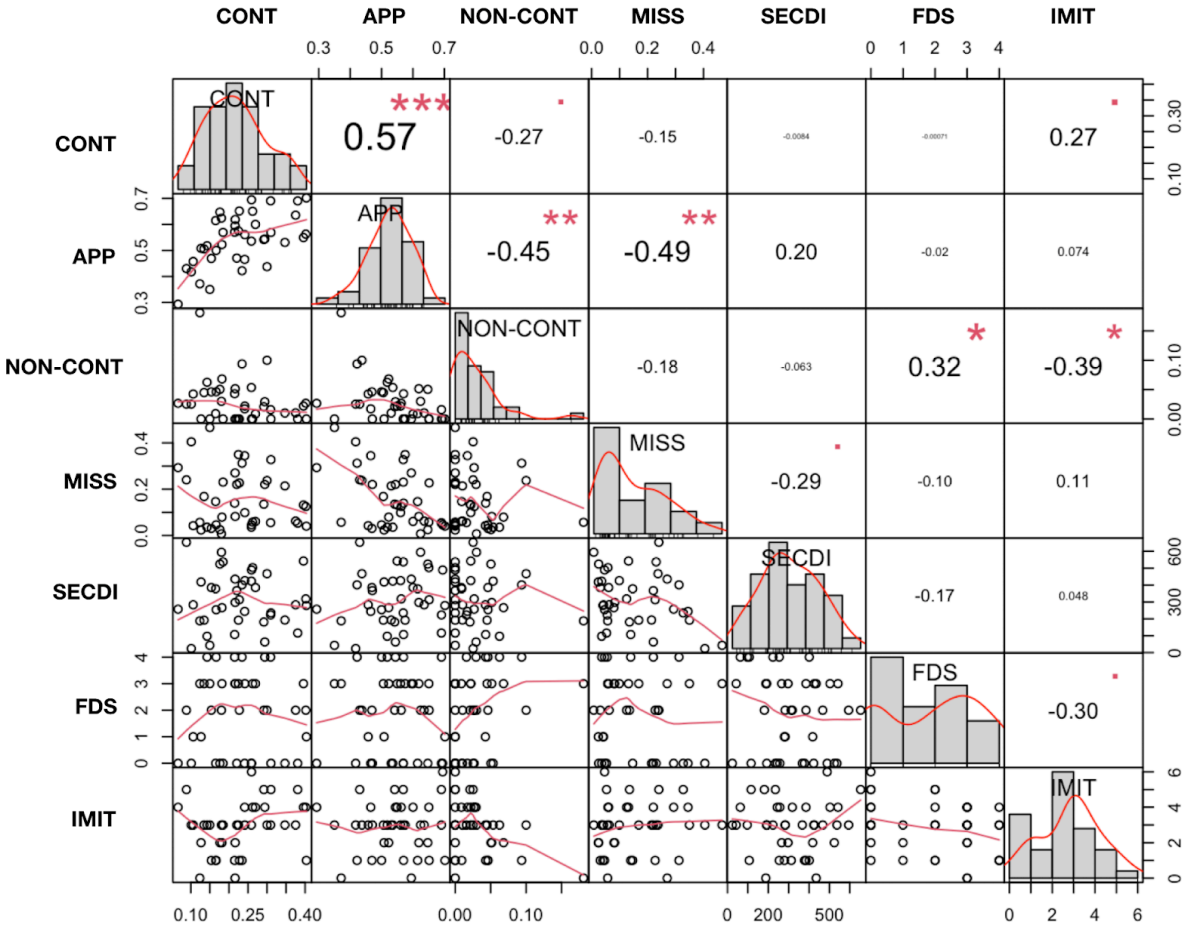


Figure 5. A correlation matrix showing pearson correlations between percentages of the four dependent variables per session: Contingent turns (CONT), Appropriate turns (APP), Non-Contingent turns (NON-CONT), Missing turns (MISS), and the predictors from Study 3 (standardized values): SECDI, forward digit span (FDS), and imitation (IMIT.).

Correlational Analyses

Figure 5 below outlines which Study 3 factors correlated with each of the four conversational measures (*Contingent turns, Appropriate turns*, *Non-Contingent turns,* and *Missing turns*) and their predictors.

Logistic regression analyses

Table 7 below reports findings for the fixed effects for each outcome variable in the logistic regression models (N = 3612), with χ^2 and p-values from the likelihood ratio test. Variance inflation factors were calculated and show no multicollinearity between predictors. Random effects for each model are presented in APPENDIX A.

Table 7. Fixed effects in Study 3 by dependent variable (Contingent turns, Appropriate turns, Non-Contingent turns, and Missing turns).

	CONTINGENT TURNS				APPROPRIATE TURNS				NON-CONTINGENT TURNS				MISSING TURNS			
	Est.	SE	χ2	p	Est.	SE	χ2	p	Est.	SE	χ2	p	Est.	SE	χ2	p
(Intercept)	-1.220	0.077	-	-	0.273	0.06	-	-	-3.86	0.198	-	-	-2.014	0.159	-	-
SECDI II	0.011	0.077	0.0206	0.886	0.106	0.06	2.941	0.086 .	0.042	0.18	0.055	0.814	-0.3	0.158	3.489	0.061 .
FDS	0.049	0.08	0.374	0.54	0.011	0.063	0.032	0.857	0.345	0.19	3.412	0.064 .	-0.077	0.165	0.218	0.64
IMITATION	0.15	0.079	3.373	0.066 .	0.032	0.063	0.265	0.606	-0.363	0.187	3.333	0.067 .	0.121	0.164	0.539	0.462

Contingent and Appropriate turns

Recall that contingent responses and appropriate responses were both positive measures of conversational ability. For both contingent responses and appropriate responses, no predictors explained significant variance in the logistic regression models (all p > .06). There is a trend towards a positive effect of early imitation for contingent turns, ((χ^2=3.37, p = .066), and of early vocabulary (SECDI) for appropriate turns, (χ^2=2.94, p = .086).

Non-contingent and missing turns

Recall that non-contingent and missing turns were both negative measures of conversational behaviour. For both non-contingent and missing turns, no predictors show reliable effects in the logistic regression models. Early imitation shows a trend towards a negative effect for *Non-Contingent turns,* (χ^2 = 3.33, p = .067) and a similar trend is found for early vocabulary (SECDI) for missing turns (χ^2 = 3.48, p = .061), Short term verbal memory (FDS) shows a trend towards a positive relationship for *Non-Contingent turns,* (χ^2 = 3.41, p = .064). Marginal R^2 for non-contingent responses was 0.072 (conditional R^2 = 0.228), the odds ratios (see figure 6) for imitation only just includes 1 [95%CI = -0.04%–48%].

Odds ratios for Study 3

W present the models in terms of odds ratios (Szumilas, 2010) with 95% confidence intervals (figure 6).

Study 3: Discussion

Early vocabulary, short-term verbal memory and early imitation showed no reliable relationships with any of the four types of conversational behaviour. In general, further exploration with a larger sample size would be needed to understand the relationship between early predictors and children's conversational behaviour.

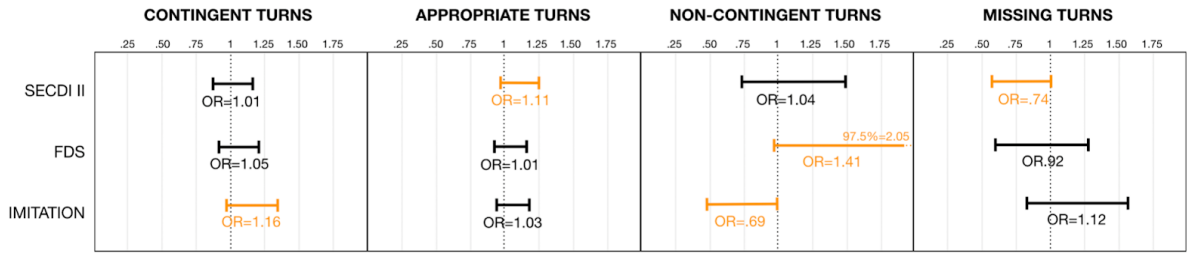


Figure 6. Odds ratios for every predictor in Study 3. The four different dependent variables are displayed in four columns (Contingent turns, Non-Contingent turns, Appropriate turns, and Missing turns). The predictors are displayed as rows. The odds ratios show how one unit in the predictor variable either increases or decreases the odds for the dependent variable to occur.

General Discussion

We carried out three studies, using one pre-existing longitudinal data set, to explore which factors might explain variance in 40 Swedish speaking 5-year-olds' conversational responses, specifically focusing on children's cognitive and social strengths in childhood, proxy measures of their environment and early measures of vocabulary, memory and imitation from infancy. In Study 1, receptive vocabulary at 4;0 predicted more *Appropriate turns,* i.e. acknowledging previous turns in general, and fewer *Missing turns,* i.e. not responding at all. In Study 2, contrary to expectation, a later age of preschool onset was associated with greater odds of responding with *Contingent turns*, i.e. responses that furthers the topic of a conversation. In Study 3, no reliable effects were found, but there were some trends that might deserve further investigation with a larger sample.

The findings from Study 1, regarding the positive relationship between receptive vocabulary at 4 years and *Appropriate turns*, aligns with previous findings - albeit with autistic children (e.g. Hale & Tager-Flusberg 2005a; Capps et al., 1998). However, to our knowledge, this is the first study to suggest a relationship between core language and a directly assessed measure of conversational ability in typically-developing children. The

predictive value of vocabulary seems easy enough to explain. If a child struggles with the comprehension of the intended meaning of lexical units in a conversation, they will struggle to understand and respond to a partner. However, it is not entirely clear why we did not find the same relationship, firstly, with (expressive) grammatical ability in Study 1 and, secondly, with parent-assessed vocabulary (Swedish version of the CDI) in Study 3 (although the latter does show a trend in the hypothesised direction, for both *Appropriate turns* and *Missing turns*). One possible explanation for the outcome of the grammatical measure would be that conversational skill is not primarily reliant on complex grammatical knowledge; someone with limited grammatical knowledge could still be appropriate and contingent, and vice versa. A non-mutually-exclusive possibility is that while an individual needs to have a certain level of morpho-syntactic ability to maintain a back-and-forth conversation, once a certain morpho-syntactic acquisition threshold has been reached, morpho-syntax no longer accounts for individual differences in child conversational ability. This possibility also helps explain the existence of a sub-group of autistic children - albeit slightly older children - who score in the high average to above-average range on morpho-syntax and vocabulary and yet find it extremely difficult to engage appropriately in reciprocal conversation (e.g. Nadig et al., 2010). Thus, future studies are required to unpack the precise relationship between lexico-grammatical knowledge, on the one hand, and appropriate responding in typically-developing children. Certainly, we assume that there is something of a two-way street between conversational development and lexical development in that it is in the context of conversation that we come to learn many words.

In our second study we investigated the role of environmental factors and found that children who started preschool later had an advantage in their *Contingent turns*. A possible explanation for this outcome is the generally high level of socioeconomic status in our sample, as well as in Sweden in general. The fact that the sample consists of families that voluntarily contributed to the longitudinal study on first language acquisition might suggest that the participating parents find language development interesting, and that they are involved in their children's development. These factors could indicate that early high quality input from a parent can aid the ability to be informative in conversation. However, preschool start did not show a reliable relationship with *Appropriate turns*, i.e. the conduct of acknowledging previous turns in general. With this in mind, and due to the complexities in measuring quality of caregiver-infant interaction and quality of daycare, this suggestion needs to be examined further, particularly in relation to possible ways in which language and conversational development could be supported in day care settings.
Finally, in Study 3, no predictors were reliably related to our four tested outcome measures. The children were very young when the imitation test was conducted, namely at 1;0, which is the developmental timepoint when fundamental abilities for understanding and sharing the basic intentions begin to be robustly evidenced (e.g. Tomasello, 2003). One way of investigating this further might be to examine imitational skill, or social cognitive insight more broadly, somewhat later in development – perhaps towards the end of the child's second year – and then assess its relationship to later conversational

ability. Future studies could also utilise imitation tasks which more closely target socio-cognitive motivation (e.g. Dohmen et al., 2013).

Limitations

While the current findings suggest avenues for future research, there are a number of limitations. Although the three studies were carried out with a rich set of available measures, the sample size was limited. The measures of conversation were reasonably ecologically valid and based on painstaking coding with excellent inter-rater reliability, but we currently do not know the test-rest reliability of this measure. When considering the short-term verbal memory measurement, it is important to note that the participants were very young and the measure might reflect knowledge of numbers more than anything else. Finally, adept conversational behavior is culturally normative and this needs to be more thoroughly explored. Studies with participants from a range of cultures will be important for understanding to what extent these results are generalizable.

Conclusion and future research

We asked which child-internal and environmental factors are related to four types of conversational behaviour when responding to an interlocutor. In line with previous findings from the literature on autistic conversation, as well as from child pragmatic development more generally –, directly-assessed receptive vocabulary was found to be a positive predictor for appropriate responding, in terms of acknowledging the turns of one's interlocutor, and a negative predictor for missing turns, i.e. not responding at all. However, neither expressive grammar nor early parent-assessed vocabulary were reliable predictors for any of the four conversational behaviours. Thus, the role of lexico-grammatical knowledge in conversational development is worth exploring further in order to understand which competencies are important limiting factors during 'live' conversation and why (e.g., due to benefits from processing speed, or depth of semantic networks, or some third variable). Contrary to what we predicted, child age when starting preschool showed a *positive* relationship with responses that further the topic of the conversation, but no reliable relationship was found with acknowledging previous turns in general. This can suggest the home environments of the children studied may have been beneficial in supporting parts of early language and communication skills (at least when observed in interaction with an adult). This needs to the explored further with respect to the quality of the home and pre-school environments. Finally, although we explored some longitudinal measures from infancy, such work would need to be done with a larger sample and better measures if one were to be certain of developmental trajectories over this time span. Overall, this preliminary exploratory study suggests an important role for lexical comprehension in responding appropriately to others. It also suggests that caregiving arrangements might influence children's conversational contingency in ways we did not initially expect, and that warrant further investigation . Future longitudinal and experimental

studies with larger sample sizes should explore the pathways that may explain such relations.

References

Abbot-Smith, K., Matthews, D., Malkin, L., Hobson, W. & Nice, J. (in preparation). Cognitive correlates of conversational contingency in autistic and neuro-typical children. DOI: 10.17605/OSF.IO/W5Y9N

Abbot-Smith, K, Matthews, D., Malkin, L. & Nice, J. (2021). On-topic conversational responding in autistic and neuro-typical children. *European Society for Philosophy and Psychology Conference, 30th August – 2nd September, Leipzig, Germany*. DOI: 10.17605/OSF.IO/Q7WA4

Ahlström, L., & Ljungman, H. (2011). Åldersreferenser för Peabody Picture Vocabulary Test IV på svenska för flerspråkiga barn i skolår 4. [Reference data for the Peabody Picture Vocabulary Test IV in Swedish for bilingual children in grade 4] (Unpublished master thesis). Karolinska Institute, Stockholm, Sweden.

Bates D, Mächler M, Bolker B, & Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. DOI: 10.18637/jss.v067.i01.

Berenguer Forner, C., Roselló, B., Baixauli Fortea, I., García Castellar, R., Colomer Diago, C., & Miranda, A. (2017). ADHD Symptoms and peer problems: Mediation of executive function and theory of mind. *Psicothema 2017, Vol. 29, No. 4*, 514-519. DOI: 10.7334/psicothema2016.376.

Bernard, S., & Deleau, M. (2007). Conversational perspective-taking and false belief attribution: A longitudinal study. British *Journal of Developmental Psychology, 25*(3), 443–460. DOI: 10.1348/026151006X171451

Berglund, E. & Eriksson, M. (2000). Communicative development in Swedish children 16-28 months old: The Swedish early communicative development inventory - words and sentences. *Scandinavian Journal of Psychology, 41*, 133-144. DOI: 10.1111/1467-9450.00181

Bishop, D. V., & Adams, C. (1989). Conversational characteristics of children with semantic-pragmatic disorder. II: What features lead to a judgement of inappropriacy? *International journal of language & communication disorders*, 24(3), 241-263. DOI: 10.3109/13682828909019890

Blain-Brière, B., Bouchard, C., & Bigras, N. (2014). The role of executive functions in the pragmatic skills of children age 4–5. *Frontiers in psychology, 5,* 240. DOI: 10.3389/fpsyg.2014.00240

Blakey, E., Visser, I., & Carroll, D. J. (2016). Different executive functions support different kinds of cognitive flexibility: Evidence from 2-, 3-, and 4-year-olds. *Child development*, *87*(2), 513-526. DOI: 10.1111/cdev.12468

Bloom, L., Rocissano, L., & Hood, L. (1976). Adult-Child Discourse: Developmental Interaction between Information Processing and Linguistic Knowledge. *Cognitive Psychology*(8), 521-552. DOI: 10.1016/0010-0285(76)90017-7

Capps, L., Kehres, J., & Sigman, M. (1998). Conversational abilities among children with autism and children with developmental delays. *Autism, 2*(4), 325-344. DOI: 10.1177/1362361398024002

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, i-174. DOI: 10.2307/1166214

Carpenter, M., Tomasello, M., & Striano, T. (2005). Role reversal imitation and language in typically developing infants and children with autism. *Infancy*, *8*(3), 253-278. DOI: 10.1207/s15327078in0803_4

Chiat, S., & Roy, P. (2013). Early predictors of language and social communication impairments at ages 9–11 years: A follow-up study of early-referred children. *Journal of Speech, Language, and Hearing Research*. *56*(6), 1824–1836. DOI: 10.1044/1092-4388

Clearfield, M. W., & Niman, L. C. (2012). SES affects infant cognitive flexibility. *Infant Behavior and Development, 35*(1), 29-35. DOI: 10.1016/j.infbeh.2011.09.007

Conti-Ramsden, G., Hutcheson, G. D., & Grove, J. (1995). Contingency and Breakdown : children with SLI and their conversations with mothers and fathers. Journal of speech and hearing research, 38, 1290 - 1302. DOI: 10.1044/jshr.3806.1290

Croft, S., Stride, C., Maughan, B., & Rowe, R. (2015). Validity of the strengths and difficulties questionnaire in preschool-aged children. *Pediatrics, 135*(5), e1210-e1219. DOI: 10.1542/peds.2014-2920

Dohmen, A., Chiat, S., & Roy, P. (2013). Nonverbal imitation skills in children with specific language delay. *Research In Developmental Disabilities, 34*(10), 3288-3300. DOI: 10.1016/j.ridd.2013.06.004

Donno, R., Parker, G., Gilmour, J., & Skuse, D. H. (2010). Social communication deficits in disruptive primary-school children. *The British Journal of Psychiatry, 196*(4), 282-289. DOI: 10.1192/bjp.bp.108.061341

Dunn, L. M., & Dunn, D. M. (2007). PPVT-4: Peabody picture vocabulary test (4th Ed.). Bloomington, MN: NCS Pearson, Inc. DOI: 10.1007/978-1-4419-1698-3_531

Farrant, B. M., Maybery, M. T., & Fletcher, J. (2011). Socio-emotional engagement, joint attention, imitation, and conversation skill: Analysis in typical development and specific language impairment. *First language*, *31*(1), 23-46. DOI: 10.1177/0142723710365431

Forman, D. R., Aksan, N., & Kochanska, G. (2004). Toddlers' responsive imitation predicts preschool-age conscience. *Psychological Science, 15*(10), 699-704. DOI: 10.1111/j.0956-7976.2004.00743.x

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental psychology, 33*(6), 966–979. DOI: 10.1037//0012-1649.33.6.966

Goodman R. (1997) The strengths and difficulties questionnaire: a research note. *Journal of Child Psychol Psychiatry. 38*(5):581–586. DOI: 10.1111/j.1469-7610.1997.tb01545.x

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58). Brill. DOI: 10.1163/9789004368811_003

Hazen, N. L., & Black, B. (1989). Preschool peer communication skills: The role of social status and interaction context. *Child development*, 867-876. DOI: 10.2307/1131028

Hale, C., & Tager-Flusberg, H. (2005a). Social communication in children with autism: The relationship between theory of mind and discourse development. *Autism, 9*, 157. DOI: 10.1177/1362361305051395

Hale, C. M., & Tager-Flusberg, H. (2005b). Brief report: The relationship between discourse deficits and autism symptomatology. *Journal of Autism and Developmental Disorders*, *35*(4), 519-524. DOI: 10.1007/s10803-005-5065-4

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Helland, W. A., Lundervold, A. J., Heimann, M., & Posserud, M. B. (2014). Stable associations between behavioral problems and language impairments across childhood–The

importance of pragmatic language problems. *Research in Developmental Disabilities, 35*(5), 943-951. DOI: 10.1016/j.ridd.2014.02.016

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368-1378. DOI: 10.1111/1467-8624.00612

Hoff-Ginsberg, E. (1987). Topic relations in mother-child conversation. *First Language*, 7(20), 145-158. DOI: 10.1177/014272378700702006

Hoff-Ginsberg, E. (1991). Mother-child conversation in different social classes and communicative settings. *Child development*, *62*(4), 782-796. DOI: 10.2307/1131177

Hoff-Ginsberg, E. (1998). The relation of birth order and socio-economic status to children's language experience and language development. *Applied Psycholinguistics*, 19, 603 - 629. DOI: 10.1017/S0142716400010389

Hughes, C. (2011). *Social understanding and social lives: From toddlerhood through to the transition to school*. Psychology Press.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*(4), 343-365. DOI: 10.1016/j.cogpsych.2010.08.002

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2325-2335. DOI: 10.1098/rstb.2009.0045

Kemple, K., Speranza, H., & Hazen, N. (1992). Cohesive discourse and peer acceptance: Longitudinal relations in the preschool years. *Merrill-Palmer Quarterly (1982-)*, 364-381.

Küntay, A. C., & Şenay, İ. (2003). Narratives beget narratives: Rounds of stories in Turkish preschool conversations. *Journal of Pragmatics, 35*(4), 559-587. DOI: 10.1016/S0378-2166(02)00129-7

Labov, W. (1969). *A study of the non-standard English of Negro and Puerto Rican speakers in New York City: cooperative research project no. 3288*. Columbia University.

Larsson, A. (2014). Barns språkutveckling: Validering av SECDI-III mot CCC-2. (Unpublished bachelor's thesis). University of Gävle, Gävle, Sweden.

Law, J., Rush, R., & McBean, K. (2014). The relative roles played by structural and pragmatic language skills in relation to behaviour in a population of primary school children from socially disadvantaged backgrounds. *Emotional and behavioural difficulties, 19*(1), 28-40. DOI: 10.1080/13632752.2013.854960

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, 44(7), 1585–1595. DOI: 10.1016/j.paid.2008.01.014

Mackie, L., & Law, J. (2010). Pragmatic language and the child with emotional/behavioural difficulties (EBD): a pilot study exploring the interaction between behaviour and communication disability. International journal of language & communication disorders, 45(4), 397-410. DOI: 10.3109/13682820903105137

Marcos, H., Salazar Orvig, A., Bernicot, J., Guidetti, M., Hudelot, C., & Préneron, C. (2004). Apprendre à parler : influence du mode de garde. Paris: L'Harmattan.

Marton, K. (2009). Imitation of body postures and hand movements in children with specific language impairment. Journal of Experimental Child Psychology, 102(1), 1-13. DOI: 10.1016/j.jecp.2008.07.007

Matthews, D., Biney, H., & Abbot-Smith, K. (2018). Individual differences in children's pragmatic ability: a review of associations with formal language, social cognition, and executive functions. *Language Learning and Development, 14*(3), 186-223. DOI: 10.1080/15475441.2018.1455584

McGillion, M., Pine, J., Herbert, J., & Matthews, D. (2017) A randomised controlled trial to test the effect of promoting caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. *Journal of Child Psychology and Psychiatry, 58*(10), 1122-113. DOI: 10.1111/jcpp.12725

Meltzoff, A. N., & Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 491-500. DOI: 10.1098/rstb.2002.1261

Nadig, A., Lee, I., Singh, L., Bosshart, K., & Ozonoff, S. (2010). How does the topic of conversation affect verbal exchange and eye gaze? A comparison between typical development and high-functioning autism. *Neuropsychologia*, *48*(9), 2730-2739. DOI: 10.1016/j.neuropsychologia.2010.05.020

Nagy, E. (2006). From Imitation to Conversation: The First Dialogues with Human Neonates. *Infant and Child Development, 15*(3), 223–232. DOI: 10.1002/icd.460

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. Journal of The Royal Society Interface, 14(134), 20170213. DOI: 10.1098/rsif.2017.0213

NICHD Early Child Care Research Network. (1999). Child care and mother-child interaction in the first 3 years of life. *Developmental Psychology*, *35*(6), 1399-1413. DOI: 10.1037/0012-1649.35.6.1399

Perner, J., Ruffman, T., & Leekam, S. (1994). Theory of Mind is contagious: you catch it from your sibs. *Child Development, 65*(4): 1228-1234. DOI: 10.2307/1131316

Place, K. S., & Becker, J. A. (1991). The influence of pragmatic competence on the likeability of grade-school children. *Discourse Processes*, *14*(2), 227-241. DOI: 10.1080/01638539109544783

Pine, J. M. (1995). Variation in vocabulary development as a function of birth order. *Child Development, 66*(1), 272-281. DOI: 10.2307/1131205

Piotrowski, J. T., Litman, J. A., & Valkenburg, P. (2014). Measuring epistemic curiosity in young children. *Infant and Child Development, 23*(5), 542-553. DOI: 10.1002/icd.1847

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2; 6 to 21 years of age. *Journal of Speech, Language, and Hearing Research, 58*(2), 345-359. DOI: 10.1044/2015_JSLHR-L-14-0150

Richardson J. T. (2007). Measures of short-term memory: a historical review. *Cortex; a journal devoted to the study of the nervous system and behavior*, *43*(5), 635–650. DOI: 10.1016/s0010-9452(08)70493-3

Rosenthal, E. N., Riccio, C. A., Gsanger, K, & M.,Pizzitola Jarratt, M. (2006) Digit Span components as predictors of attention problems and executive functioning in children. *Archives of Clinical Neuropsychology, Volume 21, Issue 2*. 131–139. DOI: 10.1016/j.acn.2005.08.004

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*(5), 1762-1774. DOI: 10.1111/j.1467-8624.2012.01805.x

Sakkalou, E., Ellis-Davies, K., Fowler, N. C., Hilbrink, E. E., & Gattis, M. (2013). Infants show stability of goal-directed imitation. *Journal of Experimental Child Psychology, 114*(1), 1-9. DOI: 10.1016/j.jecp.2012.09.005

Schulze, C., & Saalbach, H. (2021). Socio-cognitive engagement (but not socioeconomic status) predicts preschool children's language and pragmatic abilities. *Journal of Child Language,* 1-11. DOI:10.1017/S0305000921000295

Slomkowski, C., & Dunn, J. (1996). Young children's understanding of other people's beliefs and feelings and their connected communication with friends. *Developmental psychology*, *32*(3), 442. DOI: 10.1037/0012-1649.32.3.442

Song, S., Su, M., Kang, C., Liu, H., Zhang, Y., et al. (2015). Tracing children's vocabulary development from preschool through the school-age years: an 8-year longitudinal study. *Developmental Science*, 18, 119– 131. DOI: 10.1111/desc.12190

Tager-Flusberg, H., & Anderson, M. (1991). The development of contingent discourse ability in autistic children. *Journal of child psychology and psychiatry*, *32*(7), 1123-1134. DOI: 10.1111/j.1469-7610.1991.tb00353.x

Thornton, E., Matthews, D., Patalay, P., & Bannard, C. (2021, August 13). Tracking the relation between different dimensions of socio-economic circumstance and vocabulary across developmental and historical time. DOI: 10.31234/osf.io/bu3px

Tomasello, M., Conti-Ramsden, G., & Ewert, B. (1990). Young children's conversations with their mothers and fathers: differences in breakdown and repair. *Journal of Child Language*, 17, 115-130. DOI: 10.1017/S0305000900013131

Tonér, S., & Nilsson Gerholm, T. (2021). Links between language and executive functions in Swedish preschool children: A pilot study. *Applied Psycholinguistics, 42*(1), 207-241. DOI: 10.1017/S0142716420000703

Urm, A., & Tulviste, T. (2016). Sources of individual variation in Estonian toddlers' expressive vocabulary. First Language, 36(6), 580-600. DOI: 10.1177/0142723716673951

Wanska, S. K., & Bedrosian, J. L. (1985). Conversational structure and topic performance in mother-child interaction. Journal of speech and hearing research, 24, 579 - 584. DOI: 10.1044/jshr.2804.579

Zambrana, I. M., Ystrom, E., Schjølberg, S., & Pons, F. (2013). Action imitation at 1½ years is better than pointing gesture in predicting late development of language production at

3 years of age. *Child development*, *84*(2), 560-573. DOI: 10.1111/j.1467-8624.2012.01872.x

Data and script availability statement

The anonymized datasets (CC_data_210129.csv and CC_log_data_210104.csv) and the R-script for the statistical analyses (CC_full-script_210607.R) are available at https://osf.io/ah23m/.

Ethics approvals

The collection of the data used in these studies were conducted within the MINT project, The project was conducted in accordance with the regulations of The Swedish Data Protection Authority and The Ethical Review Board at Karolinska Institutet (Dnr 2011/955-31/1) and The Personal Data Act (1998:204) and The Act concerning the Ethical Review of Research Involving Humans (2003:460).

Author contribution statement

David Pagmar and Danielle Matthews devised the studies. Kirsten Abbot-Smith, Danielle Matthews, and David Pagmar outlined the rationale for the individual studies, as well as the theoretical basis for individual predictors. David Pagmar carried out the study specific data collection, coded the material, analysed the data and wrote the first draft. All authors contributed to the finalisation of the manuscript and approved the final version.

Acknowledgements

We would like to thank the families who participated, the MINT team for painstaking annotation of the interaction data, Caroline Arvidsson for additional annotation work, Linnea Rask for help with reliabilities, Thomas Hörberg for statistical advice, and Tove Gerholm for helpful comments on the manuscript. All data in this study is part of the MINT corpus, funded by the Marcus and Amalia Wallenberg Foundation (MAW2011.007) and The Swedish Research Council (2018-01135). DM was supported by British Academy grant MD\170025.

APPENDIX A

The variance and standard deviation of the random effects for each full model in Study 1, 2, and 3 in Table A.

Table A. Random effects, based on groupings by participant, in terms of variance and standard effects for the full models in studies 1 (S1), 2 (S2), and 3 (S3).

Model	Variance	SD
S1_contingent	.17	.41
S1_appropriate	.05	.24
S1_non-contingent	.87	.93
S1_missing	.84	.91
S2_contingent	.13	.36
S2_appropriate	.10	.31
S2_non-contingent	.93	.96
S2_missing	.85	.92
S3_contingent	.16	.40
S3_appropriate	.09	.31
S3_non-contingent	.66	.81
S3_missing	.87	.93

APPENDIX B

For all the three studies, each predictor's contribution to the model was evaluated through a likelihood ratio test. Although this analysis was not conducted for model selection, we present the comparative results in terms of AIC values for each run in the

likelihood ratio test. The models in Study 1 are presented in Table B1, the models in Study 2 are presented in Table B2, and the models in Study 3 are presented in Table B3.

Table B1. AIC values from the likelihood ratio test for the models on Study 1. Presented are AIC values for the full model, and for each run with one predictor excluded. The models were compared to estimate the contribution of each predictor: curiosity (CUR), the strength and difficulties questionnaire (SDQ), grammar (GRAM), and receptive vocabulary (PPVT).

Evaluation Run	AIC for Contingent model	AIC for Appropriate model	AIC for Non-contingent model	AIC for Missing model
Full model	3927.5	4878.6	966.92	2624.3
–CUR	3925.6	4876.9	964.95	2622.4
–SDQ	3925.6	4876.6	964.93	2622.5
–GRAM	3926.4	4877.6	965.07	2623.8
–PPVT	3926.6	4889.9	968.00	2626.8

Table B2. AIC values from the likelihood ratio test for the models on Study 2. Presented are AIC values for the full model, and for each run with one predictor excluded. The models were compared to estimate the contribution of each predictor: older sibling (OLD_SIB), preschool hours per week (PR.SCHO_H), age at preschool start in weeks (PR.SCHO_W), and the measure for socioeconomic status (SES).

Evaluation Run	AIC for Contingent model	AIC for Appropriate model	AIC for Non-contingent model	AIC for Missing model
Full model	3920.2	4892.0	967.81	2624.7
–OLD_SIB	3919.6	4890.3	965.91	2623.4
–PR.SCHO_H	3918.9	4891.2	965.83	2623.0
–PR.SCHO_W	3926.9	4891.3	968.18	2625.0
–SES	3918.2	4890.2	966.02	2623.3

Table B3. AIC values from the likelihood ratio test for the models on Study 3. Presented are AIC values for the full model, and for each run with one predictor excluded. The models were compared to estimate the contribution of each predictor: imitation (IMIT), forward digit span (FDS), and parental reported productive vocabulary (SECDI).

Evaluation Run	AIC for Contingent model	AIC for Appropriate model	AIC for Non-contingent model	AIC for Missing model
Full model	3924.3	4889.1	959.74	2623.3
–IMIT	3925.7	4887.4	961.15	2621.9
–FDS	3922.7	4887.2	961.07	2621.6
–SECDI	3922.3	4890.1	957.79	2624.8

License

Development of complex syntax in the narratives of children with English as an Additional Language and their monolingual peers

Disa Witkowska
Laura Lucas
Maria Jelen
Hannah Kin
Psychology and Language Sciences, University College London, UK

Courtenay Norbury
Psychology and Language Sciences, University College London, UK
Department of Special Needs Education, University of Oslo, Norway

Abstract: English syntax acquisition is crucial for developing literacy but may be challenging for many children learning English as an Additional Language (EAL). This study longitudinally investigates syntactic complexity and diversity of stories retold by children with EAL and their monolingual peers as well as the relationship between syntax and vocabulary. This is a secondary data analysis using data from the Surrey Communication and Language in Education study (SCALES). Sixty-one children with EAL were matched to their monolingual peers on sex, age and teacher-rated language proficiency. Children's narratives were collected in Year 1 (age 5-6) and Year 3 (age 7-8) and coded for clause type. Dependent variables included Mean Length of Utterance in words (MLUw) and Clausal Density (CD) as measures of syntactic complexity and Complex Syntax Type-Token Ratio (CS-TTR) estimating syntactic diversity. Children with EAL presented syntactically complex and diverse narratives equivalent to monolingual peers in Year 1 and Year 3. Growth rate in syntactic complexity was associated with English vocabulary in Year 1. Among children with low vocabulary, children with EAL developed syntactic complexity at a faster rate than monolingual peers, while the opposite was true in the high-vocabulary group. Children with average vocabulary progressed at parallel rates. Children with EAL and their monolingual peers used broadly the same complex structures but with varying frequency. In this longitudinal study comparing children with EAL and monolinguals on complex clauses, the interaction between emerging bilingualism and vocabulary knowledge in the societal language predicted different patterns of growth in syntactic complexity. Children with EAL frequently use different syntactic structures to achieve similar syntactic complexity and diversity. These findings demonstrate that in early primary school, children with EAL have syntactic skills comparable to their monolingual peers.

Keywords: bilingualism; EAL; syntactic development; complex syntax; grammar; narrative.

Corresponding author: Disa Witkowska, Division of Psychology and Language Sciences, UCL, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK. Email: disa.witkowska.15@ucl.ac.uk.

ORCID ID(s): Disa Witkowska https://orcid.org/0000-0002-8197-7301,
Laura Lucas https://orcid.org/0000-0002-9470-5284,
Maria Jelen https://orcid.org/0000-0002-9729-1208,
Courtenay Norbury https://orcid.org/0000-0002-5101-6120

Citation: Witkowska, D., Lucas, L., Jelen, M., Kin, H., & Norbury, C. (2022). Development of complex syntax in the narratives of children with English as an Additional Language and their monolingual peers. *Language Development Research, 2*(1), 180–222. https://doi.org/10.34842/2022.0551

Introduction

Worldwide, it is estimated that more people are now bi- or multilingual than monolingual (Grosjean, 2010b). In many countries, bilingual populations have increased because of immigration, which impacts on the proportion of school-age children mastering more than one language (OECD, 2019). In England, over 20 per cent of primary school pupils speak a language other than English at home (Department for Education, 2021), with implications for managing the English-dominant classroom. Limited evidence suggests that children learning English as an Additional Language (EAL) may find grammar challenging to learn (e.g. Babayiğit, 2014; Bowyer-Crane et al., 2017), but trajectories of grammar development in longitudinal cohorts have rarely compared monolinguals and those with EAL. In this paper, we track the development of complex syntax during primary school in narratives of children with EAL and their monolingual peers.

A Note on Terminology

Overlapping and sometimes inconsistent terminology, together with multiple labels used in different countries makes it difficult to define bilingualism. Broadly speaking, individuals can be considered bilingual even if the proficiency in their languages differs, if they acquired them at different ages and if they use them for different purposes (Grosjean, 2010a; Stow & Dodd, 2003). For consistency, in this paper, we use the UK education policy term "English as an Additional Language (EAL)" to describe both the study participants without making any assumptions about their home languages' proficiency and the population of children that speak more than one language. When we use an abbreviation "L2", we refer to the language of school instruction, which in this study is English.

Grammar Development in Children with EAL

Language is essential for school success and therefore for societal participation: proficiency in the language of school instruction at school entry is positively correlated with academic attainment in monolinguals (Norbury et al., 2017) and children with EAL (Whiteside et al., 2017), whose proficiency in the language of instruction covers the full spectrum of ability (Hutchinson, 2018; Strand et al., 2015).

Grammar is a key component of academic language and reading comprehension (Hjetland et al., 2020; Lervåg et al., 2018; Muter et al., 2004). The importance of grammar is recognised in the National Curriculum in England (Department for Education, 2013), which sets specific grammar targets of increasing complexity for every year

group. However, the paucity of research on grammatical development of children with EAL presents challenges in providing suitable support through education or intervention.

While the importance of vocabulary for school success has been well-established, the importance of grammar has received less research attention. A recent systematic review of language intervention studies concerning children with EAL (published between 2014 and March 2017) found that all 25 included studies featured a vocabulary component, but none targeted complex grammar (Oxley & de Cat, 2019). Given that there is a strong relationship between vocabulary development and syntactic growth in monolingual children (E. Bates & Goodman, 1997) and children with EAL (Conboy & Thal, 2006), early English vocabulary knowledge may be associated with the rate of development of complex sentences in children with EAL.

Grammar is made up of two domains: morphology, focused on the internal word structure, and syntax, concerned with the sentence structure. While a recent meta-analysis (Bratlie et al., 2022) identified morphological knowledge as a challenge for children with EAL, there is emerging evidence that syntax might be a relative strength (Paradis et al., 2017). When studies feature a single grammatical outcome conflating both domains into morphosyntax, demonstrating developmental trajectories within each domain is difficult. Our study will provide insight specifically into growth in productive syntax.

Our study can also contribute to the debate about the role of age in bilingual acquisition of grammar (see Paradis et al., 2017). The early age hypothesis posits that younger children have an advantage in learning grammar, and therefore predicts more mature English grammar for monolinguals than children with EAL of the same age. The complexity hypothesis proposes that the parallel development of language and cognitive maturity in first-language acquisition may result in protracted learning of grammar. In this case, older and cognitively mature children with EAL may need less exposure time than monolinguals to develop equivalent levels of complex English grammars.

Narrative as a Vehicle for Showcasing Syntactic Growth

Language can be sampled from naturalistic interaction, or narrative and expository tasks. The benefit of narrative is that the target is clear, relies less on the language competencies of interlocutors, and more closely resembles book language, which tends to employ more sophisticated grammar (Cameron-Faulkner & Noble, 2013; Montag, 2019). Narrative compels children to simultaneously incorporate linguistic,

cognitive and social skills to construct a logical sequence of events (Norbury & Bishop, 2003).

Narratives have been widely used in bilingualism research, in part because they are thought to be less biased than standardised tests (Boerma et al., 2016; Cleave et al., 2010). Both story generation and retelling have been used with children with EAL. Limited available evidence (see Otwinowska et al., 2020) is mixed as to whether retelling yields improved story structure and grammatical complexity in monolinguals and children with EAL. However, Otwinowska et al. (2020) showed a positive effect of retelling relative to story generation on story structure and comprehension, mental state terms and story length, but no increase in Mean Length of Utterance for both monolinguals and children with EAL.

Common methods of measuring complex syntax in narratives are presented in Table 1. Frizelle et al. (2018) used Mean Length of Utterance in words (MLUw) and Clausal Density (CD) to provide a comprehensive, cross-sectional account of the development of syntactic complexity in 354 monolingual English speakers from school entry to adulthood, using both story generation and story retell tasks. The most common clause type across all ages was the main clause, but its use decreased with age while clausal density increased with age. All clause types were present in four-year-olds' narratives, though most constructions were produced by relatively few children.

Development of Complex Clauses in Children with EAL

Monolingual English-speaking children usually start producing complex sentences after their second birthday, but the proportion of complex sentences in relation to total utterances is small until the age of four (Diessel, 2004). Complex sentences emerge type-by-type, with (non-finite) complements being first (e.g. *I wanna go,* then *I think it's a ball*), and coordinated (e.g. *I have this and you have that*), adverbial (e.g. *You can't have this cause I'm using it*) and relative clauses (e.g. *This is the toy I am playing with*) following later (Diessel, 2004).

Studies using standardised assessments of expressive grammar (e.g. sentence recall and picture description) have reported that children with EAL lag behind monolinguals in their L2 grammar (Babayiğit, 2014; Bowyer-Crane et al., 2017). However, Dixon and colleagues (2020) found no difference between the two groups, which was attributed to sufficient English language exposure prior to school entry in the EAL group.

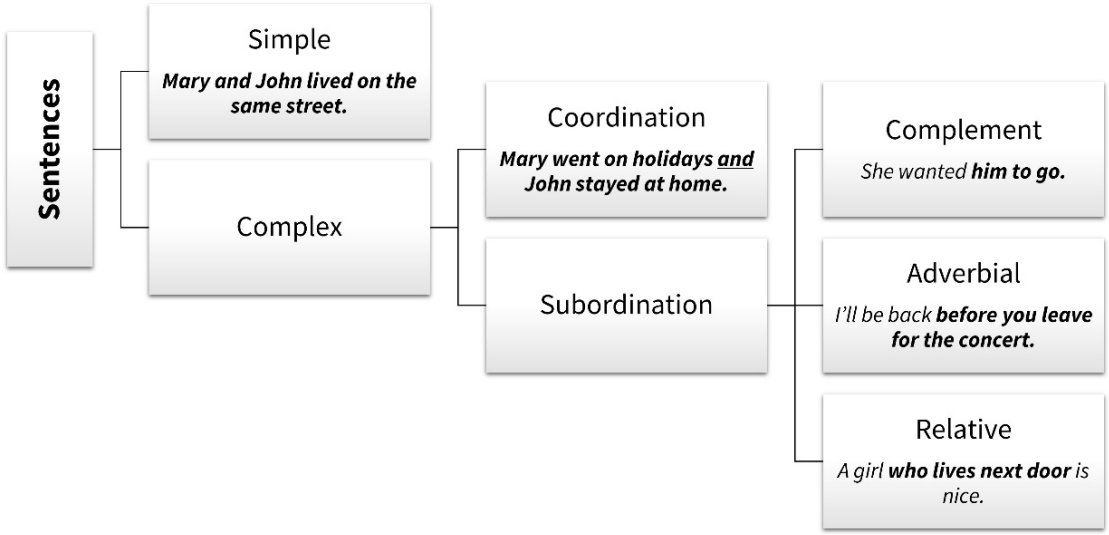


Figure 1. ***Basic classification of sentence types with examples.***

MLUw is frequently used as a measure of syntactic complexity in studies with children with EAL (e.g. Bedore et al., 2020; Simon-Cereijido & Gutiérrez-Clellen, 2009). A few studies that compared monolinguals and children with EAL (Bonifacci et al., 2018; Otwinowska et al., 2020; Rodina, 2017) produced conflicting findings, likely due to varying sample sizes (from n = 16 to 75 in groups with children with EAL), age ranges (3;1 to 7;3) and assessed languages (Norwegian, Russian, Italian and Polish).

In terms of complex syntax, Paradis et al. (2017) showed that five-year-old children with EAL needed less than a year of English exposure to start using a wide range of complex clauses without any apparent order of clause emergence. However, this study lacked an age- or language-matched monolingual comparison group, so it is unclear whether the pattern or growth in syntax is similar to that of monolinguals. Bonifacci et al. (2018) found that 4-5-year-old children with Italian as an additional language and their monolingual Italian-speaking peers produced stories with the same number of coordinate and subordinate clauses and the same proportion of complex clauses. Castilla-Earls et al. (2019) tracked the development of narrative abilities in both languages of Spanish-English speaking children using MLUw and clausal density at six points between ages 5;6 to 8;1. While English MLUw gradually increased over time, change in CD was relatively small. Children did not use subordination at all at 5;6 (CD = 1.0) and it remained minimal at four middle timepoints, peaking at 8;1 with 1.3 complex clauses per utterance. The lack of a monolingual comparison group means it is unclear whether monolinguals of that age would produce a greater

quantity or variety of syntactic structures. Additionally, children were given different stories to retell at different timepoints, which may have influenced use of clauses at any given point.

To our knowledge, only one study has directly compared clausal density (defined as the number of finite and marked infinitive clauses per utterance) of monolinguals and children with EAL (Cahill et al., 2020). This small (n < 13 in each group), cross-sectional study reported no differences between English monolinguals and English-French-speaking children in the 7-8 and 11-12-year-old group, though the authors noted that small sample size and high within-group variability limit firm conclusions. In addition, we are not aware of any longitudinal studies that have tested the extent to which proficiency in other aspects of L2 such as vocabulary may be associated with expressive syntax growth in children learning EAL.

The Current Study

We adapted the syntactic complexity framework designed by Frizelle and colleagues (2018) to investigate developmental change in syntactic complexity in a longitudinal study of children with EAL and monolingual peers from Year 1 (ages 5-6) to Year 3 (ages 7-8), using a narrative retell task. This allowed us to ask:

1. Do the narratives of children with EAL differ in syntactic complexity (MLUw and CD) from the narratives of monolingual English-speaking children in Year 1 (age 5-6) and Year 3 (age 7-8)?
2. Is the rate of growth in syntactic complexity comparable between children with EAL and their monolingual peers between Year 1 and Year 3?
3. Does English Vocabulary in Year 1 affect the rate of growth in syntactic complexity in children with EAL and their monolingual peers?
4. Do the narratives of children with EAL differ in syntactic diversity measured by Complex Syntax Type-Token Ratio (CS-TTR) relative to narratives of monolingual English-speaking children in Year 1 and Year 3?

While all children started formal schooling at the same time, children with EAL were expected to have reduced L2 syntactic complexity compared to their monolingual peers, because of reduced exposure to English language outside school. Heritage language and literacy skills can positively influence L2 acquisition, but some children with EAL still need additional English exposure to gain sufficient proficiency in English to succeed in school (see Hoff, 2013 for an overview).

The existing evidence regarding syntactic growth in children with EAL suggests that

they may develop L2 skills faster than their monolingual peers (Lonigan et al., 2013; McKean et al., 2015; Whiteside & Norbury, 2017). For example, Whiteside and Norbury showed accelerated rates of growth relative to monolingual peers between ages 5-6 and 7-8 in children with EAL on receptive vocabulary, sentence recall and overall language. This was true of children with both high and low levels of teacher-rated English language proficiency at school entry.

The strong positive relationship between the development of vocabulary and the development of grammar has been observed in early language acquisition in monolinguals (E. Bates & Goodman, 1997) and children with EAL (Conboy & Thal, 2006). A natural prediction would be to assume that better vocabulary would contribute to better grammar, hence a faster growth in syntactic complexity. However, little is known about initial vocabulary as a predictor of later syntactic growth in children with EAL, especially in comparison with monolingual peers. Conboy and Thal (2006) showed that toddlers with EAL who experienced the most growth in English vocabulary also showed the fastest rate of development of syntactic complexity in English, but demonstrated lower syntactic complexity scores at the last time point than children with slower language growth. Therefore, we tested the prediction that with increasing English vocabulary in Year 1, the rate of growth in syntactic complexity might decrease. Our study provides a strong test as we included children with a wide range of proficiency scores at school entry. EAL and monolingual groups were matched on teacher-rated English proficiency level, which ensures equal distribution of children with varying language skills across the two groups. Our longitudinal and within-subjects design featuring the same task at both time points allows us to minimise the impact of task changes and participant effects on growth estimates.

Finally, we predict that children with EAL will use fewer complex constructions than their monolingual peers, but the types of structures will be comparable across groups.

Methods

Study Design

This study is a secondary analysis using data on children with EAL and their monolingual peers from the Surrey Communication and Language in Education Study (SCALES; Norbury, Gooch, Wray, et al., 2016; Norbury, Gooch, Baird, et al., 2016; Norbury et al., 2017). First, a brief overview of the overall SCALES design is provided together with features relevant to the current study. Then follows a description of matching design and participants in the current study.

All Reception children (age 4-5) in Surrey state-maintained schools in September 2011 were invited to take part (n = 12, 398). Teachers completed questionnaires, including the Children's Communication Checklist-Short (CCC-S; see below), for 7,267 children (59% of invited children). 782 pupils (11%) spoke a language other than English at home (lower proportion than the national average in primary schools in England at that time, 16.8%) (Department for Education, 2011).

The CCC-S (Norbury et al., 2004), based on CCC-2 (Bishop, 2003a) featured seven items about communicative strengths and six about communicative errors, with higher scores (max. 39) suggesting lower English skills. Depending on CCC-S scores, three strata were identified: (1) children reported by teachers to have "no phrase speech (NPS)", based on the CCC-S item that indicates the child combines words into phrases less than once a week (assigned a maximum score), (2) "high-risk (HR)" for language disorder defined as a score 1SD or more above (indicating greater impairment) the monolingual population mean for their age group (autumn, spring, or summer born) and sex, and (3) "low-risk (LR)" for language disorder (scoring no more than 1SD above the mean for age group and sex). In this context, the term "risk" reflects teacher-reported scores on the CCC-S.

SCALES was designed to investigate individual differences in language, but not EAL per se. However, we did sample ~10% of the EAL cohort to reflect the population at the time. We included all children with no-phrase speech, and a random sample of children in the 'high-risk' group (teacher ratings of low English language proficiency relative to age and sex) and the 'low-risk' group (teacher ratings of English language proficiency in the expected range for age and sex). In this cohort, 'risk' cannot be interpreted as risk for language disorder as the CCC-S is not normed on a bilingual population. Nevertheless, it has some ecological validity in estimating children's proficiency in the language of instruction after the first year in school.

636 monolingual and 82 children with EAL from mainstream schools were invited to participate in the second part of SCALES involving intensive language assessment in Year 1 (age 5-6) and Year 3 (age 7-8). All children with NPS were invited to participate; remaining children were randomly sampled from each of the three identified strata, with equal numbers of males and females selected and a higher percentage of children at 'high-risk' of language disorder invited to participate (for further details of the selection process, see Whiteside & Norbury, 2017 and Norbury et al., 2017). In Year 1, 529 monolingual children (200 LR, 290 HR, and 39 NPS) and 61 children with EAL (25 LR, 19 HR, 17 NPS) participated. In Year 3, 499 monolingual children (192 LR, 273 HR, 35 NPS) and 51 children with EAL (21 LR, 16 HR, 14 NPS) were re-assessed.

Participants in the Current Study

61 children with EAL (29 girls) were individually matched to 61 monolingual peers on sex, language risk status (LR/HR/NPS) and age at Year 1 assessment (within 2 months). In Year 3, ten children with EAL (4 LR, 3 HR, 3 NPS) and five monolingual children (2 LR, 2 HR, 1 NPS) were lost to follow-up, therefore the final sample in Year 3 included 51 children with EAL (23 girls) and 56 monolingual children (28 girls). We did not exclude participants that had lower non-verbal reasoning or a biomedical condition. This sample partially overlaps with the sample reported by Whiteside and Norbury (2017), who analysed a sub-sample of children with EAL and monolingual peers but applied different matching criteria.

All children were recruited during the Reception Year and had at least one year of exposure to English before their Year 1 assessment. Children with EAL represented many linguistic backgrounds (24 languages spoken), with Bengali, Polish and Urdu the most frequently reported languages. The data on children's home language proficiency could not be collected due to sample heterogeneity and limited available assessments or skilled assessors in the languages required.

Socio-economic status (SES) was measured with Income Deprivation Affecting Children Index (IDACI; McLennan et al., 2011) rank scale, which is an index of neighbourhood deprivation and ranges from 1 to 32,482, based on the children's home postcode. Higher values indicate more affluent neighbourhoods with proportionally fewer households receiving means-tested benefits.

Prior to the first visit, children were randomly allocated into one of six testing blocks (half-terms in the UK school year). In Year 3, the block order was reversed (children seen in block 1 in Year 1 were seen in block 6 in Year 3 and children seen in block 6 in Year 1 were seen in block 1 in Year 3). This resulted in a variable lag of 14 to 34 months between Year 1 and Year 3 assessments, allowing us to make best use of this longitudinal design with two testing points.

Ethics and Consent Procedures

The SCALES screening phase relied on an opt-out consent procedure, allowing anonymised data from teacher questionnaires to be used in the study unless parents explicitly did not agree (20 families opted out). Informed, written consent from parents or legal guardians was required for the in-depth assessment in Year 1 and 3. The SCALES project was approved by the Ethics Committee at Royal Holloway, University of London, and further research analysis of the existing data was approved by the

Research Ethics Committee at University College London (Project ID 9733/002).

Assessment Measures

Children completed a core battery of six language assessments, comprising receptive and expressive tasks. Expressive tasks included Expressive One-Word Picture Vocabulary Test (EOWPVT; Martin & Brownell, 2011a), a sentence repetition task (SASIT-32; Marinis et al., 2011) and the information score from the narrative recall task (ACE 6-11; Adams et al., 2001). Receptive tasks included Receptive One-Word Picture Vocabulary Test (ROWPVT; Martin & Brownell, 2011b), short version (40 items) of Test for the Reception of Grammar TROG-S; (TROG Bishop, 2003b) and narrative comprehension questions. Non-verbal reasoning was measured in Year 1, using the Block Design and Matrix Reasoning subtests of the Wechsler Preschool and Primary Scale of Intelligence (Third Ed., Wechsler, 2003) (for details, see Norbury et al., 2017).

English Vocabulary in Year 1 was assessed using the Receptive One-Word Picture Vocabulary Test (ROWPVT; Martin & Brownell, 2011b). Several other measures were used to characterise the EAL and monolingual groups (see Table 3). We also used three indices Mean Length of Utterance in words (MLUw), Clausal Density (CD) and Complex Syntax Type-Token Ratio (CS-TTR) as our dependent variables (see Table 1 for explanation of concepts and our pre-registration at https://doi.org/10.17605/OSF.IO/SP24Y for implementation details).

Procedures

At each assessment point, a trained researcher met the child for a two-hour session in a quiet space in the child's school. Children completed the Assessment of Comprehension and Expression (ACE-Recall) Narrative Recall task (Adams et al., 2001), which required the child to listen to a story about a monkey and a parrot, read by an English first language speaker and played over headphones. The child simultaneously followed a PowerPoint presentation on the computer screen with eight pictures depicting the story. Immediately after the listening, the researcher asked the child to retell the story while the pictures remained on the screen. After the retelling the child was asked to answer comprehension questions, which were transcribed and scored straight after the assessment. Children's narratives were recorded using a dictaphone and later transcribed by trained student research assistants.

Table 1. ***Methods of measuring complex syntax in narratives and the rationale for using them.***

Measure	Definition	Rationale
Mean Length of Utterance in words (MLUw)	The total number of words in each utterance divided by the total number of utterances.	1. A simple way of measuring syntactic complexity development because every new grammatical construction in early child's language increases the utterance length (R. W. Brown, 1973), 2. Mainly used with children's language samples but some evidence that can successfully be used with older participants, even until adolescence and adulthood (Nippold et al., 2005).
Clausal density (CD)	The mean number of clauses per utterance, where utterance is defined as a main clause with any dependent clauses (Hunt, 1965; Loban, 1976)	MLUw might not be sufficient to assess the grammar complexity: possible to produce longer simple sentences without employing more complex syntactic structures (1). *(1) Afterwards the monkey immediately showed the parrot the juicy pineapple with a green crown.* (2) *The monkey showed the parrot the pineapple, which had a green crown.* CD rewards for a higher number of dependent clauses attached to the main clause, e.g. (1) would score 1, while (2) would score 2 (two clauses within the utterance).
Complex syntax Type-Token Ratio (CS-TTR)	The novel estimate of syntactic diversity: the mean number of different dependent construction types relative to all dependent clauses produced.	CD does not change depending on whether a speaker uses the same type of a subordinate clause throughout the narrative (3), or whether they use different types (4). (3) *The monkey showed the parrot the pineapple,* ***which had a crown*** *that was green.* *(4) After the monkey returned, he showed the parrot the pineapple,* ***which had a green crown.*** MLUw and CD provide quantitative estimates, but syntactic diversity is necessary for a more qualitative description of the development of complex sentences.

Narrative Analysis

Our coding manual (Witkowska, Lucas, & Norbury, 2021; https://osf.io/wqgz9/), based on Frizelle et al. (2018), described the process of splitting and coding the narratives. We divided sentences into clauses following a general rule of no more than one verb in each line, except for no-verb clauses (zero verbs), and g*o AND do* and *go do* constructions (two verbs but treated as one: e.g. *The monkey went and searched for treasure, Go look under the curtain*) (Frizelle et al., 2018). After splitting, narratives were transferred to Microsoft Excel and saved as comma-separated values (.csv) files.

Table 2 presents clause types distinguished in the coding manual (Witkowska, Lucas, & Norbury, 2021). Grammatical errors, word omissions or substitutions were not treated as prerequisites for discounting a clause. For example, a clause *He fellen down* was coded as a main clause despite the error in the past tense of *fall*. Where two codes were possible, we chose the code that indicated the most syntactically complex sentence. For instance, if a clause could either be coded as reported speech or imperative, we chose the first option because a main clause together with that reported speech clause would form a more syntactically complex sentence (one sentence with two clauses, i.e. (main) *The monkey said* (reported speech) *"Find me some treasure!"*) than a main clause and an imperative clause (two sentences, one clause each, i.e. (main) *The monkey said* (imperative) *"Find me some treasure!"*).

We made the following adaptations to Frizelle et al.'s (2018) coding scheme:

- Introduction of causal clause (separate codes for its finite and non-finite versions), expressing a reason for an event happening with a subordinate conjunction *because* and thus crucial for a high-quality narrative production. Previously, causal adverbial clauses (e.g. *The monkey went back* ***because*** *he was tired*) were part of an adverbial category (e.g. ***When*** *the parrot came, monkey was annoyed*), while causal non-finite non-complements (e.g. *The monkey left the tree* ***to search for treasure***) were grouped together with other non-causal non-complements (e.g. *There was a monkey* ***hanging on the high branch***).
- Separate code for imperatives (e.g. *Go to the forest!*), usually expressing commands or requests, because their lack of overt subjects makes them syntactically distinct from English main sentences.
- Separate code for verb phrases (e.g. *Locked the parrot in the cage.*) to reward children for producing more fully-developed simple sentences than no-verb utterances, despite omitting the obligatory subject.
- Preserving false starts, fillers, repetitions and unfinished sentences in the transcriptions but clearly labelling them in separate lines and excluding from syntactic complexity calculations.

Table 2. ***Codes for clause types with a short definition and a typical example.***

Code	Clause type	Description	Example
x	No-verb phrase	A non-clause which does not contain a verb.	***The end.*** ***Treasure.***
m	Main	A standalone sentence, typically following subject-verb-object word order.	***The monkey locked the parrot*** *in the cage.*
m+	Main with elided subject	A clause that could be a main clause if the subject had not been elided.	*A parrot came* ***and made lots of noise.***
cf	Finite complement	A complement clause with a marked/tensed verb.	*He knew* ***that it wasn't treasure.***
cn	Non-finite complement	A complement clause containing an unmarked verb (not indicative of tense or number).	*If you want me* ***to leave the tree…***
n	Non-finite, non-complement	A clause that contains an unmarked verb (not indicative of tense or number) and is not a compulsory part of the sentence.	*There was a monkey* ***hanging on a high branch****.*
n+	Causal non-finite non-complement	A non-compulsory clause that contains an unmarked verb and has a causal meaning	*The parrot was squawking* ***to get the monkey off the tree.***
cr	Reported speech	A complement clause that consists of a direct quotation of one of the characters.	*The parrot said "****let me out."***
a	Adverbial	A clause typically specifying locational or temporal information related to the main clause.	*I won't go away* ***until you find me some treasure.***
ca	Causal Adverbial clause	A clause that contains a cause-and-effect relationship, typically specifying a hypothetical situation with its consequences.	*The monkey went to the village* ***because he was tired****.*

Code	Clause type	Description	Example
i	Imperative	A clause without an overt subject, containing an implied subject "you."	***Don't talk to me.*** ***Go to the forest.***
vp	Verb phrase	An utterance composed exclusively of a verb phrase (missing the subject).	***Locked the cage.*** ***Was hanging on the tree.***
cc	Comment Clause	A clause expressing the speaker's attitude towards the sentence.	***I think*** *he's picking up the scarf.* ***It looks like*** *the monkey is annoyed*
co	Other comment	A clause expressing a general comment unrelated to the content of the story.	***I'm not sure.*** ***That's all I remember.***
u	Unfinished utterance	An abandoned utterance that is followed by the start of a new clause.	***He's got>*** *He's taken the parrot to the treasure.*
rr	Repetition/filler/ false start	A repetition of a word or clause, sentence-initially or otherwise; the use of filler words or just the initial letter or syllable of an intended utterance (false start).	***Ummm*** *Let me out* ***(let me out).*** ***(The m)*** *the monkey said…*
ui	Unintelligible clause	An utterance where at least 20% of the words are unintelligible and cannot be transcribed.	***The parrot *** the monkey.***

Note. These codes are a mix of Frizelle et al.'s (2018) codes together with our additions. All codes are described in detail in our syntactic coding manual (Witkowska et al., 2021).

The first and second authors prepared the narratives for coding. Two trained research assistants, the third and fourth author, coded all the transcripts, blind to group (EAL vs. Monolingual). Twenty-five narratives (out of 213, 11.7%) were double-coded by the third and fourth author. All coding queries were documented in an Excel spreadsheet and responded to by the first and second author on an ad-hoc basis. Weekly coding meetings with all the authors were an opportunity to resolve difficult issues and to ask further clarification questions. Their agreement on clause codes was good

(Cohen's Kappa = 0.85, z = 55.8, p < .001), as was the agreement on the number of grammar errors in each clause (Intra-Class Correlation, ICC = 0.75, F(1256, 1257) = 6.85, p < .001). The two coders also agreed 97 per cent of the time on verbs used in each clause.

Data analysis

This study was pre-registered on the Open Science Framework (Witkowska, Lucas, Jelen, et al., 2021; https://doi.org/10.17605/OSF.IO/SP24Y). Deviations from the plan are mentioned in the Results section. Analyses were conducted in RStudio (R Core Team, 2020) and data and analysis scripts are available on the Open Science Framework (https://osf.io/cgw9j/).

Sample Size and Power Calculation

Power curves were modelled (using pwr package; Champely, 2020) for a between-group comparison (independent-samples t-test) as a function of sample size (n = 61 for each group) for three effect-sizes d = 0.3 (small), 0.5 (medium) and 0.8 (large). The modelling showed 80% power to detect an effect size of 0.5, and 38% power to detect an effect size of 0.3.

Missing Data

Narrative data were available for 54 children with EAL and 55 monolingual children in Year 1, and for 51 children with EAL and 53 monolinguals in Year 3. Children who were seen for assessment but did not produce a story (6 children with EAL and 4 monolinguals in Year 1, and 3 monolinguals in Year 3) were assigned a score of 0 on each outcome measure to reflect their minimal expressive language.

Missing narratives that were excluded from analysis included those with no audio-recording (1 child with EAL and 2 monolinguals in Year 1) and families lost to follow-up (10 children with EAL and 5 monolinguals in Year 3). Children who were not followed-up in Year 3 did not consistently differ from those who remained in the study on any of the measured variables, including socio-economic status (EAL group: $M_{no\text{-}follow\text{-}up}$ = 18124.2 and M_{rest} = 17218.2, p = .753; MONO group: $M_{no\text{-}follow\text{-}up}$ = 24344.00 and M_{rest} = 21757.38, p = .457); vocabulary Year 1 (EAL group: $M_{no\text{-}follow\text{-}up}$ = 65.5 and M_{rest} = 69.55, p = .419; MONO group: $M_{no\text{-}follow\text{-}up}$ = 75.4 and M_{rest} = 77.29, p = .8); or ACE Narrative Information scores in Year 1 (EAL group: $M_{no\text{-}follow\text{-}up}$ = 11.75 and M_{rest} = 9.72, p = .325; MONO group: $M_{no\text{-}follow\text{-}up}$ = 10.75 and M_{rest} = 10.87, p = .96).

We had intended to use Full Information Maximum Likelihood estimation to account for missing data, but this could not be used within the framework of *lme4* as pre-registered. However, one advantage of linear mixed models (LMMs) is that only an observation at a specific time point is excluded from the analysis, not all observations from the same participant, and thus LMMs are robust to handle the missing data. That allowed use of data from 60 children with EAL and 59 monolingual children in Year 1 and 51 children with EAL and 56 monolingual children in Year 3. In total, 226 observations were used in each LMM.

Statistical Analysis for Confirmatory Analyses

We employed linear mixed models (LMMs), using lme4 package (D. Bates et al., 2015), that account for the non-independence of the data (V. A. Brown, 2020), that is, the fact that within-children scores were more similar to each other than between-children scores. LMMs are also robust to unequal sample sizes (Baayen et al., 2008). We acknowledge that the growth in the measures of interest might not be linear, however, a growth curve analysis with quadratic or cubic terms could not be implemented with only two testing points.

For Research Questions 1-3, two separate LMMs with MLUw and CD as dependent variables were run, with Group (EAL vs. MONO), Age (in months) and English Vocabulary in Year 1 (ROWPVT-4 score) as fixed effects and Child ID as by-participants random intercept. The models also contained the following interactions: Group x Age, Group x English Vocabulary, Age x English Vocabulary, and Group x Age x English Vocabulary. To correctly interpret the interactions, Age and Vocabulary scores were centred, thus 0 means an average age in Year 1 and an average vocabulary score in Year 1 respectively. We used Age (in months) instead of Timepoint to account for our use of variable testing lags between each Timepoint (Year 1 and Year 3).

A maximal random effect structure (Barr et al., 2013) comprised by-participants (Child ID) random intercept to account for the initial variation in the complexity of the children's narratives. By-participants random slope was not possible because we had only one observation (one MLUw or CD score) per child per timepoint.

For Research Question 4, a separate LMM was constructed with CS-TTR as dependent variable. It included Group and Age fixed effects, by-participants (Child ID) random intercept and the Group x Age interaction.

Results

Background Measures

Children with EAL and their monolingual peers were matched on sex, age at Year 1 (within two months), and their teacher-rated, English language proficiency status (NPS/HR/LR) derived from their CCC-S score (see Table 3).

Table 3. ***Descriptive statistics for background variables for EAL and Monolingual groups (raw scores are provided for standardised assessment).***

Variable	EAL	MONO	t-test	
	M (SD)	M (SD)	*t*(df)	*p*
Year 1 Participants – n	61	61	NA	NA
Female – n (%)	29 (47.5%)	29 (47.5%)		
Year 3 Participants - n	51	56	NA	NA
Female – n (%)	23 (45%)	28 (50%)		
Year 1 Age (months)	71.34 (4.15)	71.43 (4.24)	-0.11 (120)	0.914
Year 3 Age (months)	95.45 (4.54)	94.21 (4.25)	1.46 (105)	0.148
Year 1 - Year 3 Lag (months)	24.43 (5.6)	22.84 (5.3)	1.51 (105)	0.134
CCC-S	21.43 (13.82)	19.93 (14.83)	0.57 (120)	0.567
IDACI Rank	17366.72 (8224.72)	21969.39 (7373.43)	-3.25 (120)	0.001
Non-verbal reasoning	25.8 (4.17)	25.62 (4.57)	0.23 (119)	0.815
Year 1 Receptive Vocabulary	68.89 (14.33)	77.13 (15.74)	-3.03 (120)	0.003
Year 3 Receptive Vocabulary	96.16 (14.34)	94.73 (16.99)	0.47 (104)	0.642
Year 1 Receptive Grammar	20.66 (8.84)	23.63 (7.76)	-1.97 (119)	0.051
Year 3 Receptive Grammar	26.8 (7.25)	28.98 (7.79)	-1.48 (103)	0.142
Year 1 Narrative information score	10.02 (5.34)	10.86 (4.44)	-0.91 (110)	0.365
Year 3 Narrative information score	15.82 (4.6)	14.47 (5.48)	1.36 (102)	0.177

Note. Abbreviations: CCC-S – Children's Communication Checklist-Short; SES – Socio-Economic Status operationalised as IDACI rank; Non-Verbal reasoning = Block Design and Matrix Reasoning subtests from the Wechsler Preschool and Primary Scale of Intelligence; Receptive Vocabulary = ROWVPT-4; Receptive Grammar = TROG-S; Narrative Information Score derived from the ACE Narrative sub-scale.

Children in the two groups did not differ with respect to age at Year 3, time lag between Year 1 and Year 3 assessments, or non-verbal reasoning. Children with EAL lived in more economically deprived areas and had poorer English vocabulary in Year 1, but not in Year 3, compared to monolingual peers. Receptive grammar (TROG-S) was marginally lower for the EAL group relative to monolingual pupils in Year 1, but not in Year 3. The groups did not differ on narrative information scores at either time point, indicating that their stories contained a similar number of key narrative events.

Narrative Characteristics

Prior to the main analysis, children's narratives were characterised with respect to several factors potentially relevant for the explanation of the main findings.

Table 4. ***Means and SDs of narrative characteristics between EAL and Monolingual groups in Years 1 and 3.***

Variable	EAL		MONO					
	Year 1	Year 3	Year 1	Year 3	Year 1 t-test		Year 3 t-test	
	M (SD)	M (SD)	M (SD)	M (SD)	*t*(df)	*p*	*t*(df)	*p*
Utterances (n)	17.96 (7.65)	22.94 (7.03)	16.62 (5.01)	18.98 (5.92)	1.08 (91.16)	.281	3.11 (102)	.002
Dependent clauses (n)	7.65 (5.52)	12.37 (6.7)	6.72 (4.48)	12.8 (6.6)	0.94 (100)	.348	-0.33 (100)	.744
Different verbs (n)	14.74 (6.64)	20.02 (5.26)	14.24 (4.5)	18.62 (5.67)	0.46 (93.05)	.644	1.3 (102)	.196
Grammar errors (n)	4.37 (3.19)	3.61 (3.86)	3.42 (3.63)	2.32 (1.95)	1.45 (107)	.149	2.13 (73.27)	.036
Children with at least one grammar error - n (% of all children in that group)	52 (96%)	48 (94%)	45 (82%)	45 (85%)	NA	NA	NA	NA

Note. Calculation excludes 10 children in Year 1 and 3 children in Year 3 who did not produce the narrative.

Table S1 (supplementary materials) shows the number of clause codes excluded from the main analysis (unfinished and unintelligible utterances, comments unrelated to the story, repetitions, fillers and false starts). There were numerically more repetitions and false starts in the EAL group than in the monolingual group.

Children produced stories of similar length in Year 1, while in Year 3 children with EAL produced longer stories than monolingual peers (see Table 4). Children in the two groups at both time points employed a similar number of dependent clauses. The mean number of grammar errors was numerically higher in the EAL than in the monolingual group at both time points but the difference was statistically significant only in Year 3. Of all children who produced a narrative, the vast majority committed at least one grammatical error at both time points, but the proportion of children who made at least one such error was numerically higher in the EAL group in both Years 1 and 3.

Children in the two groups used a comparable number of distinct verbs at both time points. A wider range of verbs was employed in Year 3 relative to Year 1. Figure S2 (in supplementary materials) illustrates that the top 10 most frequently employed verbs – likely driven by the narrative content – by children with EAL and their monolingual peers were almost the same, with "be", "find" and "say" always being in the top 3.

Correlations

Pearson's correlations are provided in Figure 2 as they not only show the relationships between key variables but might also be useful for future meta-research. Syntactic complexity indices were more stable between Year 1 and Year 3 in the monolingual group relative to the EAL group (see Figure 2 for Pearson's correlations), indicating more variation in growth trajectories within the EAL group relative to the Monolingual group.

Main Analysis

Research Question 1-3: Syntactic Complexity

The means and standard deviations of the outcome measures are in Table 5, while the distribution of MLUw and CD is shown in Figure 3. Contrary to the pre-registration, we decided not to exclude outliers as we were interested in children who span the range of language proficiency. Removing extreme, but relatively frequent, observations would not address the heterogeneity of language skills in both groups and therefore blur the real-life picture. As models with MLUw and CD as dependent variables

(see Table 6) had statistically significant interactions, the lower-order effects could not be interpreted as main effects but as simple effects, when all other predictors are equal to 0 (V. A. Brown, 2020).

There was no simple effect of Group for participants of average age and English vocabulary in Year 1, and no two-way interactions.

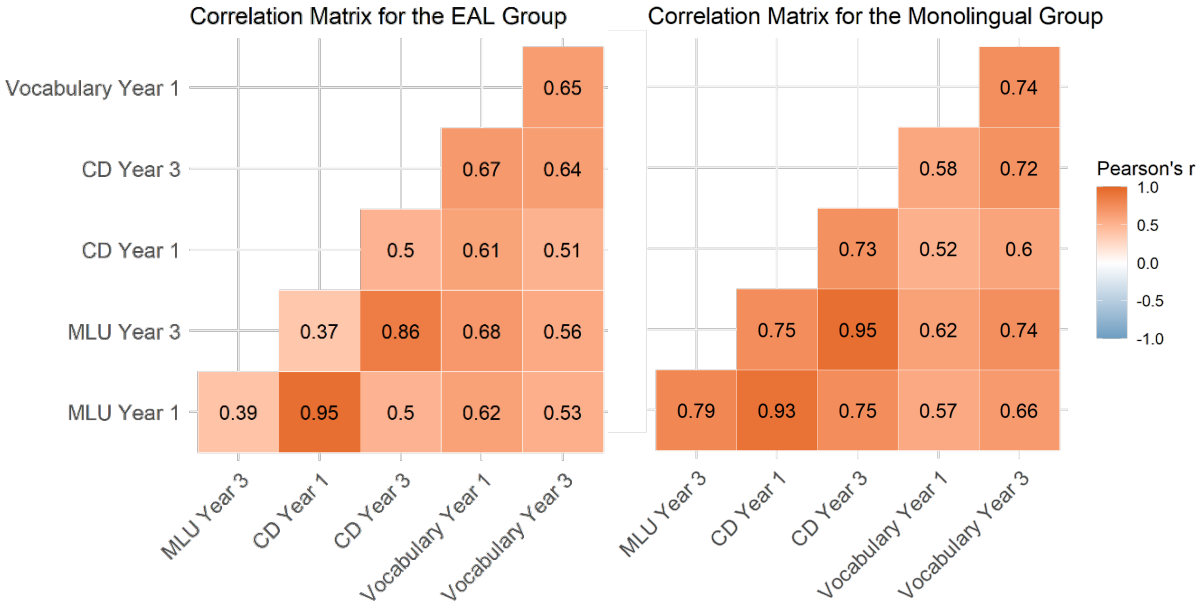


Figure 2. ***Correlations between Year 1 and Year 3 syntactic complexity indices (MLUw and CD) as well as English vocabulary for EAL and Monolingual groups. All correlations were highly statistically significant (p < .009).***

Table 5. ***Descriptive statistics for syntactic complexity (MLUw and CD) and syntactic diversity (CS-TTR) indices for EAL and Monolingual groups in Years 1 and 3.***

Outcome	EAL		MONO	
	Year 1	Year 3	Year 1	Year 3
	M (SD)	M (SD)	M (SD)	M (SD)
MLUw	5.95 (2.73)	8.02 (1.24)	6.38 (2.33)	7.82 (2.56)
CD	1.15 (0.53)	1.52 (0.28)	1.21 (0.43)	1.49 (0.5)
CS-TTR	0.54 (0.34)	0.5 (0.2)	0.55 (0.28)	0.48 (0.26)

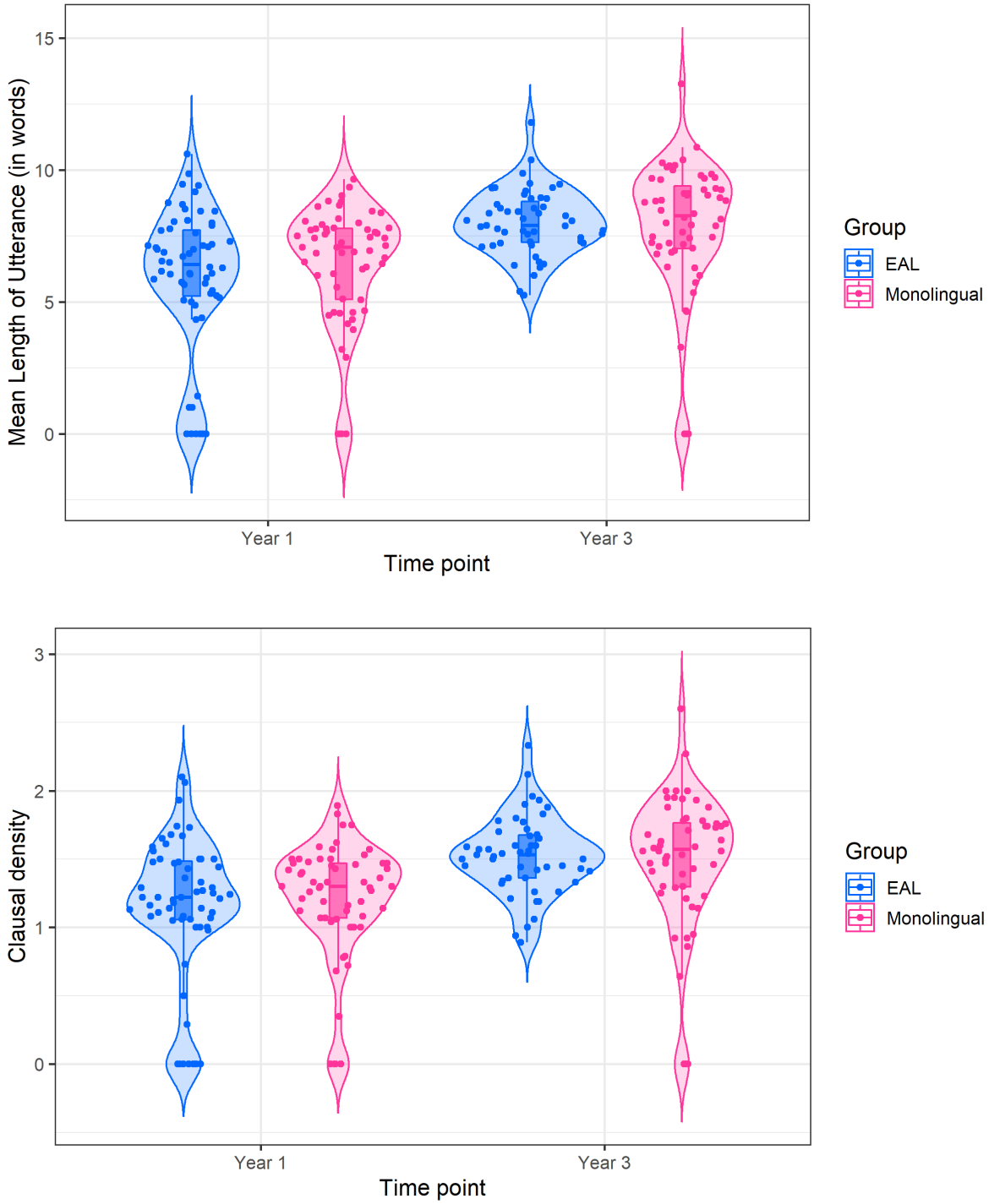


Figure 3. ***Distributions of syntactic complexity indices (MLUw and CD) for the EAL and Monolingual groups in Years 1 and 3.***

The statistically significant Group x Age x Vocabulary interaction indicated that the pattern of growth in syntactic complexity is different for EAL and monolingual groups. It is also dependent on the English vocabulary size in Year 1 (see Figure 4). For the EAL group, the higher the English vocabulary knowledge in Year 1, the lower the rate of growth in syntactic complexity. For the monolingual group, it was the opposite; the rate of syntactic growth increased with higher vocabulary size in Year 1.

Table 6a. ***Results of the linear mixed model with MLUw as a dependent variable.***

Effect	Estimate	SE	95% CI		*p*
			Lower	Upper	
Fixed					
Intercept	6.083	0.235	5.620	6.546	<.001
Group	0.390	0.335	-0.271	1.050	.246
Age	0.054	0.011	0.033	0.075	<.001
English Vocabulary	0.084	0.015	0.055	0.114	<.001
Group x Age	0.020	0.015	-0.010	0.050	.181
Group x English Vocabulary	0.038	0.022	-0.005	0.082	.083
Age x English Vocabulary	0.001	0.001	0.000	0.002	.188
Group x Age x English Vocabulary	-0.003	0.001	-0.005	-0.001	0.001

Table 6b. ***Results of the linear mixed model with CD as a dependent variable.***

Effect	Estimate	SE	95% CI		*p*
			Lower	Upper	
Fixed					
Intercept	1.149	0.047	1.056	1.242	<.001
Group	0.093	0.067	-0.040	0.225	.168
Age	0.011	0.002	0.007	0.015	<.001
English Vocabulary	0.015	0.003	0.009	0.020	<.001
Group x Age	0.003	0.003	-0.003	0.009	.280
Group x English Vocabulary	0.009	0.004	0.000	0.017	.054
Age x English Vocabulary	0.000	0.000	0.000	0.000	.102
Group x Age x English Vocabulary	-0.001	0.000	-0.001	0.000	.004

Note. SE = Standard Error, CI = Confidence Interval. Group: 0 = monolingual, Age – centred: 0 = mean age in Year 1, English Vocabulary – centred: 0 = mean vocabulary in Year 1.

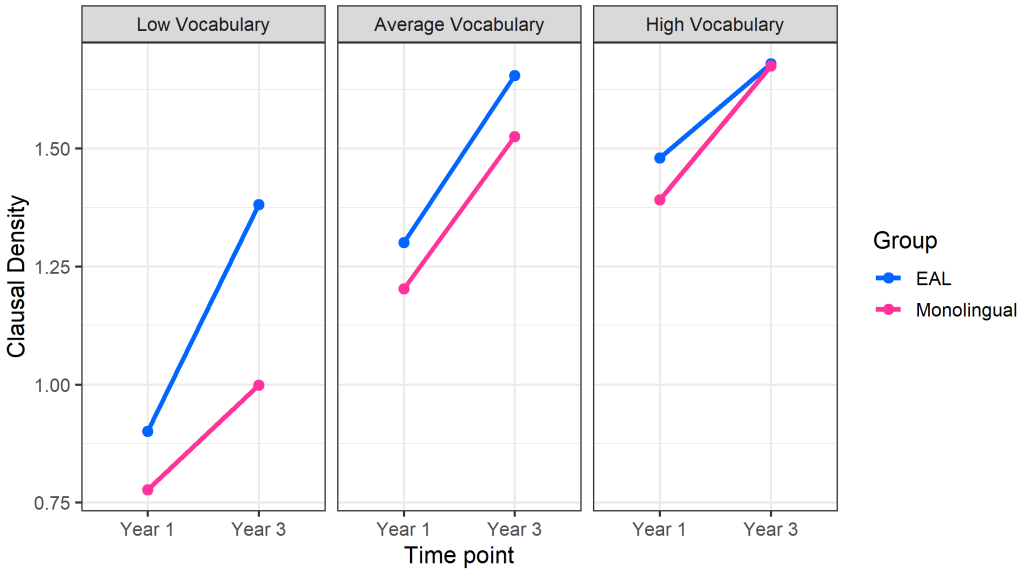


Figure 4a. ***Mean growth trajectories in CD between Year 1 and Year 3 for children with EAL and their monolingual peers with Low (below the EAL mean vocabulary score in Year 1, 68.89; n = 33 and n = 13 respectively), Average (between the EAL mean (68.89) and the monolingual mean (77.13) vocabulary score in Year 1; n = 8 and n = 17 respectively) and High (above the monolingual mean vocabulary score in Year 1; n = 20 and n = 31 respectively) English vocabulary in Year 1.***

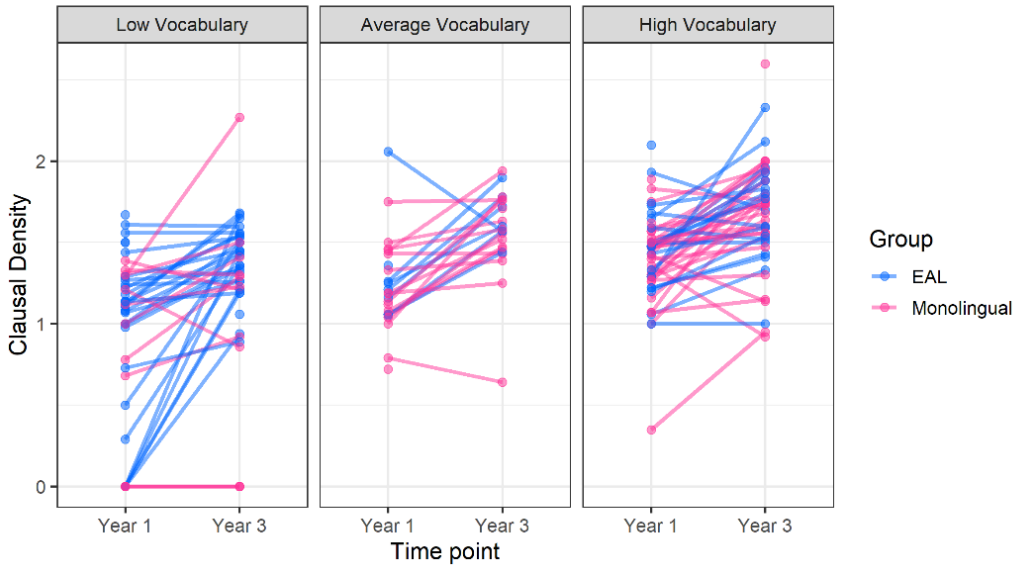


Figure 4b. ***Individual growth trajectories in CD between Year 1 and Year 3 for children with EAL and their monolingual peers with Low, Average and High English Vocabulary in Year 1.***

Children with EAL and lowest vocabulary scores (below the EAL mean vocabulary score in Year 1, 68.89) experienced faster growth in syntactic complexity than monolingual children with similar vocabulary scores. Children with EAL who had average vocabulary (between the EAL mean (68.89) and the monolingual mean (77.13) vocabulary score in Year 1) showed roughly parallel rates of growth in syntactic complexity to their monolingual peers with the same average vocabulary size. For children with EAL whose vocabulary in Year 1 was above the monolingual mean, the predicted rate of growth in syntactic complexity decreased and fell below the monolingual rate of growth, while monolingual children with the highest vocabulary scores in Year 1 exhibited the fastest growth in syntactic complexity among all monolingual participants.

Research Question 4: Syntactic Diversity

CS-TTR was introduced to quantify syntactic diversity in addition to syntactic complexity. However, many CS-TTR scores were located on the edges of the distribution, taking a value of 0, indicating that all clauses were the same, or a value of 1, showing that each clause was of a different type. The residuals distribution was not normal, therefore we could not run a linear mixed model with CS-TTR as a dependent variable.

Following Frizelle and colleagues (2018), we report the proportion of children who retold the story and produced at least one example of a given clause type (see Table 7). Almost all children could construct a main sentence, but there was a substantial proportion of children in both groups that used verb phrases or no-verb utterances. In Year 1, more than two out of five children in both groups resorted to no-verb phrases. In Year 3, this figure dropped considerably in the EAL group, but remained similar in the monolingual group.

We can also see different patterns of clause use employed by children with EAL and their monolingual peers over time. In Year 1, similar proportions of children across the two groups used finite complements (cf), relative (r) and non-complement non-finite (n) clauses. In Year 3, 69 per cent of children with EAL employed finite complements compared to 45 per cent of monolingual children. The opposite was found for non-complement non-finite and relative clauses, with a higher proportion of monolingual children using these types of clauses than children with EAL (42 vs. 33% and 47 vs. 31% respectively).

Table 7. ***Proportions of children who retold the story and produced at least one example of a given clause type for EAL and Monolingual groups in Years 1 and 3.***

Clause code	Clause type	EAL		MONO	
		Year 1	Year 3	Year 1	Year 3
m	main	0.96	1	1	1
m+	main with elided subject	0.69	0.82	0.58	0.91
cr	reported speech	0.63	0.88	0.62	0.83
cn	non-finite complement	0.57	0.78	0.55	0.75
ca	causal adverbial	0.54	0.75	0.56	0.72
x	no-verb phrase	0.41	0.24	0.45	0.4
a	adverbial	0.33	0.45	0.2	0.58
vp	verb phrase	0.31	0.2	0.31	0.26
cf	finite complement	0.3	0.69	0.36	0.45
r	relative	0.22	0.33	0.18	0.42
n	non-finite, non-complement	0.19	0.31	0.18	0.47
i	imperative	0.15	0.22	0.07	0.3
n+	causal non-finite non-complement	0.07	0.12	0.11	0.26
cc	comment clause	0.06	0.04	0.04	0.08

With respect to clauses most relevant for constructing a coherent story, different developmental patterns were observed for causal adverbials (ca) and non-complement non-finite clauses with a causal meaning (n+). Causal adverbials (e.g. *so he couldn't talk, 'cause this is not your tree, if you bring me some treasure*) were used by a similar proportion of children in both groups at both time points (above 50% in Year 1 and almost 75% in Year 3). Non-finite clauses with a causal meaning (e.g. *[then he's going out] to get some things, [so the monkey set out] to find some treasure*), produced by a smaller proportion of children, were employed by more monolingual children than children with EAL at both time points, with the difference being especially large in Year 3 (12 vs. 26%).

Table S3 (supplementary materials) demonstrates the frequency of clause use in the children's narratives. Children in both groups employed main clauses roughly two-thirds of the time in Year 1, but they became less frequent in Year 3, particularly in the monolingual group. Overall, there were no large differences between the two groups at either time point, as different types of complex clauses appeared roughly the same number of times as in the narratives of monolingual children and those with EAL.

Exploratory Analysis

The Relationship Between Growth in Syntactic Complexity and Growth in English Vocabulary

We further investigated what motivates faster growth in syntactic complexity in the EAL low-vocabulary group. Our hypothesis was that it might be related to greater growth in English vocabulary.

An additional LMM with English Vocabulary as dependent variable, Age and Group as fixed effects and by-participants random intercept estimated a significant Group x Age interaction (β = 0.319, SE = 0.092, 95% CI [0.136, 0.501], p <.001), which indicated that children with EAL indeed developed their vocabulary faster than their monolingual peers.

Then, associations between the magnitude of growth in syntactic complexity and in vocabulary were computed. Growth in vocabulary, MLUw and CD was calculated as a difference between Year 1 and 3 raw scores.

Table 8a. ***Pearson's correlations between growth in syntactic complexity (MLUw and CD) and English vocabulary for all participants with observations at both time points (n = 106).***

Correlation	EAL		MONO	
	r	*p*	*r*	*p*
MLUw growth – Vocabulary growth	0.05	.733	0.09	.514
CD growth – Vocabulary growth	0.12	.397	0.18	.201

Table 8b. ***Pearson's correlations between growth in syntactic complexity (MLUw and CD) and English vocabulary for participants whose growth on each variable was greater than 0 (n = 76).***

Correlation	EAL		MONO	
	r	*p*	*r*	*p*
MLUw growth – Vocabulary growth	0.24	.136	<0.001	.984
CD growth – Vocabulary growth	0.22	.177	0.17	.321

Pearson's correlations between growth in vocabulary and growth in syntactic complexity (both MLUw and CD) were weak and not statistically significant in both EAL and monolingual groups (see Table 8), despite moderate-to-strong correlations between syntactic complexity (MLUw and CD) and English vocabulary in Years 1 and 3 (see Figure 2).

The same correlations were calculated for 76 out of 106 children who exhibited positive growth on each outcome measure to account for regression to the mean and measurement errors. Correlations in the monolingual group were even weaker than previously, while correlations in the EAL group were considerably larger, although did not reach statistical significance.

The Effect of SES on Syntactic Complexity

Given group differences in SES, we included SES as a covariate in the LMMs with MLUw and CD as dependent variables. SES had a statistically significant effect on both measures (MLUw: $\beta = 0.081$, SE = 0.031, 95% CI [0.021, 0.142], p =.009; CD: $\beta = 0.017$, SE = 0.007, 95% CI [0.004, 0.03], p =.012). In neither case did the inclusion of SES alter the main finding, that level of Year 1 English Vocabulary is associated with growth in syntactic complexity. However, we note that due to the sample size, the models did not have sufficient power to examine a four-way interaction.

Discussion

Our study follows a unique cohort over a two-year period that spans a range of English language proficiency, has been in formal English language schools from school entry, and has been measured on the same narrative assessment on two occasions. This gives us a rare opportunity to look at the development of complex syntax using a more naturalistic task. Having matched children with EAL and their monolingual peers for English language proficiency at school entry, we see rather few differences between groups on syntactic complexity or growth. However, early levels of English vocabulary may differentially influence the rate of growth in syntactic complexity in the two groups. What is also note-worthy is the rapid progress of children with EAL at the tail of the Year 1 distribution, which could reflect their increased exposure to rich academic language. We now consider our research questions in more detail.

Did the Narratives of Children with EAL and Their Monolingual Peers Differ in Syntactic Complexity?

Contrary to our predictions, we found no difference in syntactic complexity (MLUw

and CD) in Year 1 and Year 3 between the narratives of children with EAL and their monolingual peers. Mean syntactic complexity scores in our study were broadly similar to previous reports (Cahill et al., 2020; Castilla-Earls et al., 2019; Frizelle et al., 2018), although including children with varying English language skills in our study resulted in more variation than in the previous studies.

Our results provide stronger evidence for Cahill et al.'s (2020) report of no statistically significant difference between children with EAL and their monolingual peers on syntactic complexity. Our sample was also more linguistically diverse, thus the finding can be extended beyond French-English speaking children in the unique Canadian environment. Most importantly, our study is a longitudinal study and therefore gives more direct evidence for developmental trajectories than previous cross-sectional work.

There are several potential reasons for the similarities in syntactic complexity between children with EAL and their monolingual peers. First, we assumed that children in the EAL group may have had less exposure to English at home, but we could not verify that assumption. Thus, children with EAL could have had English exposure comparable to their monolingual counterparts, or at least sufficient exposure to produce stories of similar syntactic complexity. Dixon and colleagues (2020) found that most children with EAL were born in the UK and received substantial English input at home, which – they argued – might have attenuated group differences in their sample. Furthermore, one-year exposure to English during the first school year may increase exposure to academic language, which includes more complex grammatical forms than conversational English (Snow & Uccelli, 2009).

In addition, the quantity of input may be less important than the 'readiness' of children to make use of that input (see Paradis et al., 2017). The complexity hypothesis proposes that since cognitive maturity develops at the same time as language skills in first language learners, it can restrict the frequent use of complex constructions. This limitation would not apply to children L2 learners, as they would be older and thus more cognitively mature when exposed to L2, and therefore they could start producing complex clauses after a shorter language exposure than their monolingual counterparts. This could explain why children from the low-vocabulary EAL group, who had average non-verbal reasoning, were able to use school input to accelerate their language learning. In turn, slow growth in syntactic complexity in the low-vocabulary monolingual group might reflect reduced language input but could also be indicative of broader neurodevelopmental difficulties (such as language disorder) that make it more challenging to learn language from typical home or school input.

Our matching design meant that children with different levels of teacher-rated English language proficiency at school entry were distributed evenly across the EAL and monolingual groups. Considering the heterogeneity of language skills in both groups enabled us to estimate the effect of bilingualism, without confounding it with initial differences in English language proficiency. As a side note, our design might have contributed not only to similar syntactic complexity in the two groups, but also to the EAL group "catching up" in receptive vocabulary by age 7-8, an unusual finding in the literature (e.g. compare with Dixon et al., 2022). Very few studies employ such matching; usually a random sample of children with EAL and monolingual peers is selected, in contrast to our more balanced sample. This means that in our study children with EAL did not have to aim that high to achieve results comparable to their monolingual peers.

Furthermore, the narrative retelling task might have constrained the range of syntactic structures produced, enhancing similarities between the groups. The narratives exhibited striking similarities in both groups (e.g. equal story length, frequent use of the same verbs) and exposure to the model story might have provided useful (or necessary) scaffolding, enabling children with EAL to demonstrate their best storytelling and syntactic skills. This scaffolding may be less important for monolinguals, especially those with good vocabulary knowledge.

Similar syntactic complexity in the narratives of children with EAL and their monolingual peers also offers an interesting insight into the distinction between two components of grammar: syntax and morphology. Most children in the two groups at both time points committed at least one grammatical error and children with EAL committed more grammatical errors than their monolingual peers. Although we did not code specific error types, syntactic errors (such as wrong word order) were rare, whereas morphological errors were common (e.g. missing 3rd person singular *–s*, or past tense *–ed*). This would indicate that morphology might be a relative weakness of children with EAL (Bratlie et al., 2022), while complex sentences are a relative strength (Paradis et al., 2017).

Finally, we were unable to assess grammatical complexity in the child's home language(s) but acknowledge that this might play a role in the development of English syntax. Grammatical features can transfer from one language to another (Yip & Matthews, 2007), which might be responsible for ungrammatical or atypical constructions (Otwinowska et al., 2020). Simultaneously, there is some evidence that hearing a syntactic construction in one language can make children with EAL more likely to produce this construction in another language (e.g. Hervé et al., 2016; Vasilyeva et al., 2010; Wolleb et al., 2018), even if the primed construction is ungrammatical in the

target language (Hsin et al., 2013). This suggests that a heritage language can provide scaffolding for children to learn similar constructions in another language, which could compensate, at least to some extent, for lesser exposure to the societal language.

Did the EAL and Monolingual Groups Differ in the Rate of Growth in Syntactic Complexity?

In general, both groups experienced growth in syntactic complexity during the two-year period. However, growth trajectories for the EAL and monolingual groups depended on the English vocabulary knowledge in Year 1. Among children with low English vocabulary in Year 1, syntactic complexity developed faster in the EAL group relative to monolingual peers, but the opposite was true among children with high vocabulary. Children with average vocabulary showed parallel rates of growth irrespective of whether they spoke EAL.

Notably, most children with EAL with poorer English language skills experienced rapid growth in syntactic complexity over the first three years in school, consistent with the complexity hypothesis. In contrast, monolingual children with low language skills demonstrated slower rates of growth that may indicate more general issues with language learning (Whiteside & Norbury, 2017). The slower growth in complex syntax of the high-vocabulary children with EAL than for the high-vocabulary monolinguals is quite surprising but suggestive of regression to the mean.

Overall, these findings add to the existing evidence that early proficiency in the language of instruction better predicts language growth and outcomes than the EAL label alone (Hessel & Strand, 2021; Whiteside & Norbury, 2017).

Despite moderate-to-strong associations between syntactic complexity (MLUw and CD) and English vocabulary at both time points, vocabulary growth was not correlated with growth in syntactic complexity in neither group. This seems to be consistent with Valentini and Serratrice's (2021) finding that in children with EAL in early primary school, vocabulary and grammar develop independently. Together with results of correlated growth in these two domains in younger children with EAL (aged 2;6 to 4; Hoff et al., 2018), it appears likely that there are developmental effects in the relationship between growth in vocabulary and growth in grammar. Our exploratory finding is thus worth replicating on in future studies with more assessment points.

Did the Narratives of the EAL and Monolingual Groups Differ in Syntactic Diversity?

In addition to the frequency with which complex syntax was produced, we were also interested in the range of syntactic forms that children included in their narratives. Children with EAL used a similar range of constructions to their monolingual peers, but some types of complex clauses were produced with varying frequency in the two groups. All construction types were present in both groups in Year 1 but increased in use to Year 3.

In sum, children with EAL were able to construct narratives with comparable number of utterances and clauses as their monolingual peers, and their stories were equally complex, although this was achieved through using different types of clauses with different frequency. Our findings provide evidence that bilinguals are not two monolinguals in one (Grosjean, 1989), as children with EAL in our study displayed different, but not detrimental, trajectories of syntactic diversity development.

Strengths and Limitations of the Study

This study has many strengths: it is one of few longitudinal studies comparing syntactic complexity of children with EAL and their monolingual peers over a two-year period. Using a population sample, we employed a matching design ensuring that children with different levels of English language skills were evenly distributed across the two groups. Our participants with EAL were from linguistically-diverse backgrounds, which is the more typical situation in community schools (as opposed to a single language community). Finally, our reliable and detailed coding manual could be used by educators to track the types of constructions used by children with EAL and mapped to grammatical forms targeted in the National Curriculum.

Our study is limited by the lack of data on home language exposure, both concurrent and prior to school entry. This would have allowed us to compare the English input in the monolingual and EAL groups and quantify the extent of the possible cross-linguistic transfer. However, in the UK context with over 300 languages spoken in schools (NALDIC, 2012), it is difficult for schools to collect this type of information about their pupils, and there is a lack of reliable assessment and qualified assessors to obtain such information directly. Additionally, despite a relatively large sample size giving us enough power to detect effect sizes of 0.5 or more, we had less power to detect smaller differences between the EAL and monolingual groups. Yet, the numerically higher Year 3 syntactic complexity in the EAL group than in monolinguals indicates the unexpected direction of the effect, which could be replicated in future

studies with larger sample sizes.

Our groups differed with respect to socio-economic disadvantage, despite recruitment from a generally more affluent area of the UK. Inclusion of SES as a co-variate did not affect our primary findings, but the potentially different role that SES may play for children with and without EAL on language development requires further investigation with larger samples and more diverse socio-economic backgrounds.

Furthermore, our linear mixed models were able to account for initial language ability differences across children (random intercepts) but could not take into consideration by-participant differences in the rate of change. To construct models with random slopes, a longitudinal study with at least three time points is necessary.

The study also spotlighted one caveat to using a narrative task despite its many benefits: children might produce stories that are not a true reflection of their underlying maximal language skills. Therefore, replicating the analysis of the relationship between vocabulary and growth in syntactic complexity using different tasks (for example, expository discourse) would be necessary to examine the consistency of the effects we found in this study.

Educational Implications

Our results can serve as reference data on the development of complex sentences in children with EAL and their monolingual peers. Furthermore, story retelling appears to be a useful pedagogical tool for assessing children's knowledge of syntactic constructions and identifying practice targets, minimising word-finding demands for the EAL group.

Conclusions

We found no difference between children with EAL and their monolingual peers on syntactic complexity, but different developmental patterns of syntactic diversity. Growth in syntactic complexity varied by initial English vocabulary knowledge, with the fastest growth experienced by low-vocabulary children with EAL and high-vocabulary monolingual children. Children with EAL made more grammatical errors than monolinguals at both time points but achieved comparable syntactic complexity, which suggests that errors might create a false perception of their relatively strong syntactic skills.

References

Adams, C., Cooke, R., Crutchley, A., Hesketh, A., & Reeves, D. (2001). *Assessment of Comprehension and Expression 6-11 (ACE 6-11).* GL assessment.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Babayiğit, S. (2014). The role of oral language skills in reading and listening comprehension of text: A comparison of monolingual (L1) and bilingual (L2) speakers of English language. *Journal of Research in Reading, 37*(S1), S22–S47. https://doi.org/10.1111/j.1467-9817.2012.01538.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, E., & Goodman, J. C. (1997). On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. *Language and Cognitive Processes, 12*(5–6), 507–584. https://doi.org/10.1080/016909697386628

Bedore, L. M., Peña, E. D., Fiestas, C., & Lugo-Neris, M. J. (2020). Language and Literacy Together: Supporting Grammatical Development in Dual Language Learners With Risk for Language and Learning Difficulties. *Language, Speech, and Hearing Services in Schools, 51*(2), 282–297. https://doi.org/10.1044/2020_LSHSS-19-00055

Bishop, D. V. M. (2003a). *Children's Communication Checklist—Second Edition.* Pearson.

Bishop, D. V. M. (2003b). *Test for Reception of Grammar–Version 2.* Pearson Assessment.

Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: Implications for clinical practice. *International Journal of Language & Communication Disorders, 51*(6), 626–638. https://doi.org/10.1111/1460-6984.12234

Bonifacci, P., Barbieri, M., Tomassini, M., & Roch, M. (2018). In few words: Linguistic gap but adequate narrative structure in preschool bilingual children. *Journal of Child Language, 45*(1), 120–147. https://doi.org/10.1017/S0305000917000149

Bowyer-Crane, C., Fricke, S., Schaefer, B., Lervåg, A., & Hulme, C. (2017). Early literacy and comprehension skills in children learning English as an additional language

and monolingual children with language weaknesses. *Reading and Writing, 30*(4), 771–790. https://doi.org/10.1007/s11145-016-9699-8

Bratlie, S. S., Brinchmann, E. I., Melby-Lervåg, M., & Torkildsen, J. von K. (2022). Morphology—A Gateway to Advanced Language: Meta-Analysis of Morphological Knowledge in Language-Minority Children. *Review of Educational Research*, 1–37. https://doi.org/10.3102/00346543211073186

Brown, R. W. (1973). *A first language: The early stages.* Allen and Unwin.

Brown, V. A. (2020). *An introduction to linear mixed effects modeling in R*. PsyArXiv. https://doi.org/10.31234/osf.io/9vghm

Cahill, P., Cleave, P., Asp, E., Squires, B., & Bird, E. K.-R. (2020). Measuring the complex syntax of school-aged children in language sample analysis: A known-groups validation study. *International Journal of Language & Communication Disorders, 55*(5), 765–776. https://doi.org/10.1111/1460-6984.12562

Cameron-Faulkner, T., & Noble, C. (2013). A comparison of book text and Child Directed Speech. *First Language*, *33*(3), 268–279. https://doi.org/10.1177/0142723713487613

Castilla-Earls, A., Francis, D., Iglesias, A., & Davidson, K. (2019). The Impact of the Spanish-to-English Proficiency Shift on the Grammaticality of English Learners. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1739–1754. https://doi.org/10.1044/2018_JSLHR-L-18-0324

Champely, S. (2020). *Pwr: Basic Functions for Power Analysis. R package version 1.3-0.* https://CRAN.R-project.org/package=pwr

Cleave, P. L., Girolametto, L. E., Chen, X., & Johnson, C. J. (2010). Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders*, *43*(6), 511–522. https://doi.org/10.1016/j.jcomdis.2010.05.005

Conboy, B. T., & Thal, D. J. (2006). Ties Between the Lexicon and Grammar: Cross-Sectional and Longitudinal Studies of Bilingual Toddlers. *Child Development*, *77*(3), 712–735. https://doi.org/10.1111/j.1467-8624.2006.00899.x

Department for Education. (2011). *Schools, pupils and their characteristics: January 2011*. https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2011

Department for Education. (2013). *English programmes of study: Key stages 1 and 2*. https://www.gov.uk/government/publications/national-curriculum-in-england-english-programmes-of-study

Department for Education. (2021). *Schools, pupils and their characteristics: January 2021*. https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics

Diessel, H. (2004). *The Acquisition of Complex Sentences*. Cambridge University Press.

Dixon, C., Hessel, A., Smith, N., Nielsen, D., Wesierska, M., & Oxley, E. (2022). Receptive and expressive vocabulary development in children learning English as an additional language: Converging evidence from multiple datasets. *Journal of Child Language*, 1–22. https://doi.org/10.1017/S0305000922000071

Dixon, C., Thomson, J., & Fricke, S. (2020). Language and reading development in children learning English as an additional language in primary school in England. *Journal of Research in Reading*, *43*(3), 309–328. https://doi.org/10.1111/1467-9817.12305

Frizelle, P., Thompson, P. A., McDonald, D., & Bishop, D. V. M. (2018). Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, *45*(5), 1174–1197. https://doi.org/10.1017/S0305000918000144

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15. https://doi.org/10.1016/0093-934X(89)90048-5

Grosjean, F. (2010a). Describing Bilinguals. In *Bilingual: Life and Reality*. (pp. 18–27). Harvard University Press. https://www.jstor.org/stable/j.ctt13x0ft8.6

Grosjean, F. (2010b). Why Are People Bilingual? In *Bilingual: Life and Reality*. (pp. 3–17). Harvard University Press. https://www.jstor.org/stable/j.ctt13x0ft8.5

Hervé, C., Serratrice, L., & Corley, M. (2016). Dislocations in French–English bilingual children: An elicitation study. *Bilingualism: Language and Cognition*, *19*(5), 987–1000. https://doi.org/10.1017/S1366728915000401

Hessel, A. K., & Strand, S. (2021). Proficiency in English is a better predictor of educational achievement than English as an Additional Language (EAL). *Educational Review*, 1–24. https://doi.org/10.1080/00131911.2021.1949266

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. *Educational Research Review*, *30*, 100323. https://doi.org/10.1016/j.edurev.2020.100323

Hoff, E. (2013). Interpreting the Early Language Trajectories of Children From Low-SES and Language Minority Homes: Implications for Closing Achievement Gaps. *Developmental Psychology*, *49*(1), 4–14. https://doi.org/10.1037/a0027238

Hoff, E., Quinn, J. M., & Giguere, D. (2018). What explains the correlation between growth in vocabulary and grammar? New evidence from latent change score analyses of simultaneous bilingual development. *Developmental Science, 21*(2), e12536. https://doi.org/10.1111/desc.12536

Hsin, L., Legendre, G., & Omaki, A. (2013). Priming Cross-Linguistic Interference in Spanish-English Bilingual Children. *Proceedings of the 37th Annual Boston University Conference on Language Development.*

Hunt, K. W. (1965). *Grammatical structures written at three grade levels: NCTE Research Report No. 3*. National Council of Teachers of English. https://files.eric.ed.gov/fulltext/ED113735.pdf

Hutchinson, J. (2018). *Educational Outcomes of Children with English as an Additional Language*. The Education Policy Institute, The Bell Foundation and Unbound Philanthropy. https://www.bell-foundation.org.uk/eal-programme/research/educational-outcomes-of-children-with-english-as-an-additional-language/

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the Developmental Relationship Between Oral Language Skills and Reading Comprehension: It's Simple, But Complex. *Child Development, 89*(5), 1821–1838. https://doi.org/10.1111/cdev.12861

Loban, W. (1976). *Language development: Kindergarten through grade twelve. NCTE Research report No. 18*. National Council of Teachers of English. https://files.eric.ed.gov/fulltext/ED128818.pdf

Lonigan, C. J., Farver, J. M., Nakamoto, J., & Eppe, S. (2013). Developmental Trajectories of Preschool Early Literacy Skills: A Comparison of Language-Minority and Monolingual-English Children. *Developmental Psychology, 49*(10), 1943–1957. https://doi.org/10.1037/a0031408

Marinis, T., Chiat, S., Armon-Lotem, S., Piper, J., & Roy, P. (2011). *School Age Sentence Imitation Task–English 32.* Unpublished test. https://researchcentres.city.ac.uk/language-and-communication-science/veps-very-early-processing-skills/veps-assessments

Martin, N. A., & Brownell, R. (2011a). *Expressive One-Word Picture Vocabulary Test–Fourth Edition.* Academic Therapy Publications.

Martin, N. A., & Brownell, R. (2011b). *Receptive One-Word Picture Vocabulary Test–Fourth Edition.* Academic Therapy Publications.

McKean, C., Mensah, F. K., Eadie, P., Bavin, E. L., Bretherton, L., Cini, E., & Reilly, S. (2015). Levers for Language Growth: Characteristics and Predictors of Language Trajectories between 4 and 7 Years. *PLoS One, 10*(8), e0134251. https://doi.org/10.1371/journal.pone.0134251

McLennan, D., Barnes, H., Michael, N., Davies, J., Garratt, E., & Dibben, C. (2011). *The English indices of deprivation 2010*. Department for Communities and Local Government. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010-technical-report

Montag, J. L. (2019). Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language*, *39*(5), 527–546. https://doi.org/10.1177/0142723719849996

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, Rimes, Vocabulary, and Grammatical Skills as Foundations of Early Reading Development: Evidence From a Longitudinal Study. *Developmental Psychology*, *40*(5), 665–681. https://doi.org/10.1037/0012-1649.40.5.665

NALDIC. (2012). *Languages in schools: More about the languages of bilingual pupils.* https://www.naldic.org.uk/research-and-information/eal-statistics/lang/

Nippold, M. A., Hesketh, L. J., Duthie, J. K., & Mansfield, T. C. (2005). Conversational Versus Expository Discourse: A Study of Syntactic Development in Children, Adolescents, and Adults. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1048–1064. https://doi.org/10.1044/1092-4388(2005/073)

Norbury, C. F., & Bishop, D. V. M. (2003). Narrative skills of children with communication impairments. *International Journal of Language & Communication Disorders*, *38*(3), 287–313. https://doi.org/10.1080/136820310000108133

Norbury, C. F., Gooch, D., Baird, G., Charman, T., Simonoff, E., & Pickles, A. (2016). Younger children experience lower levels of language competence and academic progress in the first year of school: Evidence from a population study. *Journal of Child Psychology and Psychiatry*, *57*(1), 65–73. https://doi.org/10.1111/jcpp.12431

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry*, *57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Norbury, C. F., Nash, M., Baird, G., & Bishop, D. V. M. (2004). Using a parental checklist to identify diagnostic groups in children with communication impairment: A validation of the Children's Communication Checklist—2. *International Journal of Language & Communication Disorders*, *39*(3), 345–364. https://doi.org/10.1080/13682820410001654883

Norbury, C. F., Vamvakas, G., Gooch, D., Baird, G., Charman, T., Simonoff, E., & Pickles, A. (2017). Language growth in children with heterogeneous language disorders:

A population study. *Journal of Child Psychology and Psychiatry, 58*(10), 1092–1105. https://doi.org/10.1111/jcpp.12793

OECD. (2019). *The Road to Integration: Education, Migration and Social Cohesion*. OECD. https://doi.org/10.1787/d8ceec5d-en

Otwinowska, A., Mieszkowska, K., Białecka-Pikul, M., Opacki, M., & Haman, E. (2020). Retelling a model story improves the narratives of Polish-English bilingual children. *International Journal of Bilingual Education and Bilingualism, 23*(9), 1083–1107. https://doi.org/10.1080/13670050.2018.1434124

Oxley, E., & de Cat, C. (2019). A systematic review of language and literacy interventions in children and adolescents with English as an additional language (EAL). *The Language Learning Journal, 49*(3), 265–287. https://doi.org/10.1080/09571736.2019.1597146

Paradis, J., Rusk, B., Duncan, T. S., & Govindarajan, K. (2017). Children's Second Language Acquisition of English Complex Syntax: The Role of Age, Input, and Cognitive Factors. *Annual Review of Applied Linguistics, 37*, 148–167. https://doi.org/10.1017/S0267190517000022

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rodina, Y. (2017). Narrative abilities of preschool bilingual Norwegian-Russian children. *International Journal of Bilingualism, 21*(5), 617–635. https://doi.org/10.1177/1367006916643528

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2009). A cross-linguistic and bilingual evaluation of the interdependence between lexical and grammatical domains. *Applied Psycholinguistics, 30*(2), 315–337. https://doi.org/10.1017/S0142716409090134

Snow, C. E., & Uccelli, P. (2009). The Challenge of Academic Language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge Handbook of Literacy* (pp. 112–133). Cambridge University Press. https://doi.org/10.1017/CBO9780511609664.008

Stow, C., & Dodd, B. (2003). Providing an equitable service to bilingual children in the UK: A review. *International Journal of Language & Communication Disorders, 38*(4), 351–377. https://doi.org/10.1080/1368282031000156888

Strand, S., Malmberg, L., & Hall, J. (2015). *English as an Additional Language (EAL) and educational achievement in England: An analysis of the National Pupil Database*. The Education Endowment Foundation, Unbound Philanthropy and The Bell Foundation. https://www.bell-foundation.org.uk/app/uploads/2017/05/EALachievementStrand-1.pdf

Valentini, A., & Serratrice, L. (2021). What Can Bilingual Children Tell Us About the Developmental Relationship Between Vocabulary and Grammar? *Cognitive Science, 45*(11), e13062. https://doi.org/10.1111/cogs.13062

Vasilyeva, M., Waterfall, H., Gámez, P. B., Gómez, L. E., Bowers, E., & Shimpi, P. (2010). Cross-linguistic syntactic priming in bilingual children. *Journal of Child Language, 37*(5), 1047–1064. https://doi.org/10.1017/S0305000909990213

Wechsler, D. (2003). *Wechsler Preschool and Primary Scales of Intelligence–Third UK Edition.* Pearson Assessment.

Whiteside, K. E., Gooch, D., & Norbury, C. F. (2017). English Language Proficiency and Early School Attainment Among Children Learning English as an Additional Language. *Child Development, 88*(3), 812–827. https://doi.org/10.1111/cdev.12615

Whiteside, K. E., & Norbury, C. F. (2017). The Persistence and Functional Impact of English Language Difficulties Experienced by Children Learning English as an Additional Language and Monolingual Peers. *Journal of Speech, Language, and Hearing Research, 60*(7), 2014–2030. https://doi.org/10.1044/2017_JSLHR-L-16-0318

Witkowska, D., Lucas, L., Jelen, M., Kin, H., & Norbury, C. (2021). *Development of complex syntax in the narratives of children with English as an Additional Language and their monolingual peers [Open Science Framework pre-registration].* https://doi.org/10.17605/OSF.IO/SP24Y

Witkowska, D., Lucas, L., & Norbury, C. (2021). *Syntactic complexity coding manual.* https://doi.org/10.17605/OSF.IO/WQGZ9

Wolleb, A., Sorace, A., & Westergaard, M. (2018). Exploring the role of cognitive control in syntactic processing: Evidence from cross-language priming in bilingual children. *Linguistic Approaches to Bilingualism, 8*(5), 606–636. https://doi.org/10.1075/lab.17002.wol

Yip, V., & Matthews, S. (2007). *The Bilingual Child: Early Development and Language Contact*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620744

Data, code and materials availability statement

The coded narrative data, participants' characteristics and R scripts for data wrangling and analysis are available on Open Science Framework at https://osf.io/cgw9j/. The syntactic coding manual can be found at https://doi.org/10.17605/OSF.IO/WQGZ9.

We endeavoured to make all the relevant documentation as open as possible, however, in the following cases it was not possible, and the Editor, Ben Ambridge, approved the exemptions from sharing on 21st January 2022.

We were unable to share any data that may be identifiable. Therefore, we did not share IDACI Rank scores, which use children's home postcode to estimate neighbourhood affluence. In addition, the small number of children speaking particular languages and the narrow geographical area from which we recruited could potentially identify participants. We therefore did not include children's home language data.

Furthermore, the standardised assessments that we used (most importantly Narrative Retell task from Assessment of Comprehension and Expression, Adams et al., 2001) are copyrighted and thus it was not possible for us to openly share the test material.

Ethics statement

Parents or legal guardians gave informed, written consent for the in-depth assessments in Year 1 and 3. The SCALES project was approved by the Ethics Committee at Royal Holloway, University of London, and further research analysis of the existing data was approved by the Research Ethics Committee at University College London (Project ID 9733/002).

Authorship and Contributorship Statement

DW was involved in conceptualisation and design of the study, led on creation of the syntactic complexity coding manual, participated in preparing narrative data for coding, co-supervised narrative data coding, analysed the data, prepared figures and tables, wrote the first draft and edited subsequent drafts. LL was involved in creating the syntactic complexity coding manual, took the lead on narrative data curation and preparation for coding as well as planning of narrative coding, co-supervised narrative data coding, contributed to writing Assessment Measures and Procedures sections of the paper, and reviewed and edited the manuscript draft. MJ and HK gave feedback on the syntactic complexity coding manual, coded the narrative data, prepared the table of clause types in the Methods section, and reviewed and edited the manuscript draft. CN conceptualised and designed the study, was responsible for overseeing research activity planning and execution throughout the study, provided supervision, and reviewed and edited the manuscript draft.

All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This research was supported by the Economic and Social Research Council Studentship (ES/P000592/1) awarded to Disa Witkowska and the Wellcome Trust (WT094836AIA) grant awarded to Courtenay Norbury.

We would like to thank Jo Saul and Sarah Griffiths for ongoing R coding support and helpful discussion of the findings. We would like to thank Aleksandra Petrykiewicz for assistance with writing the code for matching participants. We also thank Suet Yee Ng and Lik Hong Lo for starting the work at adapting Frizelle et al.'s (2018) coding manual for our purposes. Finally, we thank all the schools, parents and children who participated in this research project.

Supplementary Materials

Table S1. ***The mean number of other comments (co), repetitions (false starts and fillers; rr), unfinished (u) and unintelligible (ui) utterances for the EAL and Monolingual groups in Years 1 and 3.***

Clause type	EAL		MONO		Comparison between EAL and MONO			
	Year 1	Year 3	Year 1	Year 3	Year 1 t-test		Year 3 t-test	
	M (SD)	M (SD)	M (SD)	M (SD)	*t*(df)	*p*	*t*(df)	*p*
co	2.73 (2.88)	2.46 (2.35)	2.58 (1.84)	2.5 (2)	0.26 (48.47)	.799	-0.06 (52)	.952
rr	11.63 (8.96)	11.69 (6.49)	8.67 (6.04)	10.44 (7.2)	2.02 (92.96)	.047	0.92 (101)	.36
u	2.32 (1.75)	2.08 (1.4)	2.14 (1.24)	1.84 (1.07)	0.42 (56)	.677	0.84 (72)	.404
ui	1.5 (0.55)	1.83 (2.04)	1.25 (0.5)	1.81 (2.26)	0.73 (8)	.486	0.02 (20)	.984

Note. Calculation excludes 10 children in Year 1 and 3 children in Year 3 who did not produce the narrative.

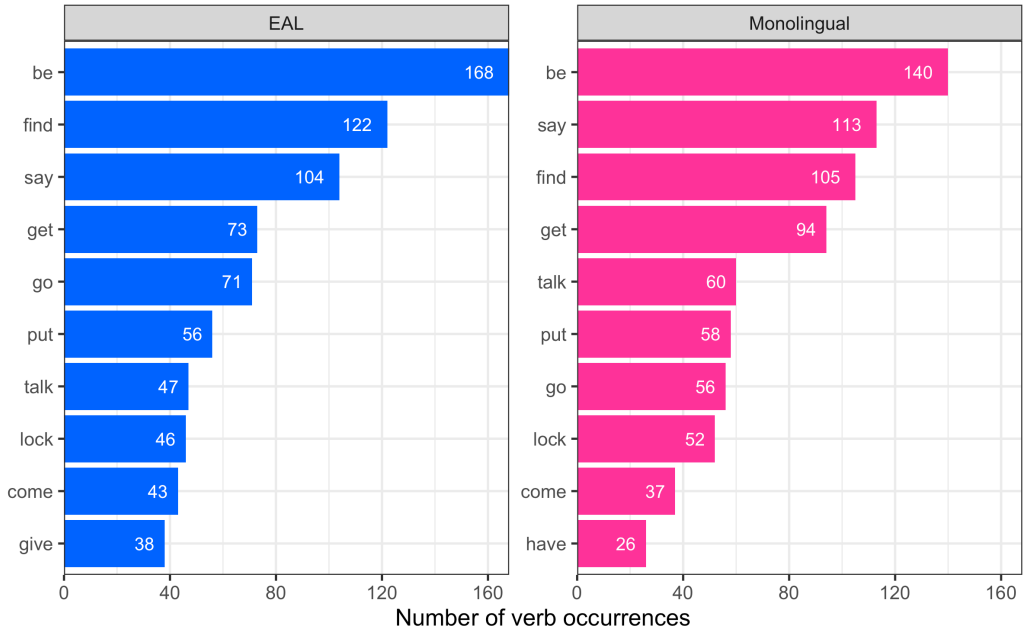


Figure S2a. ***Comparison between EAL and Monolingual groups on 10 most frequent verbs in Year 1.***

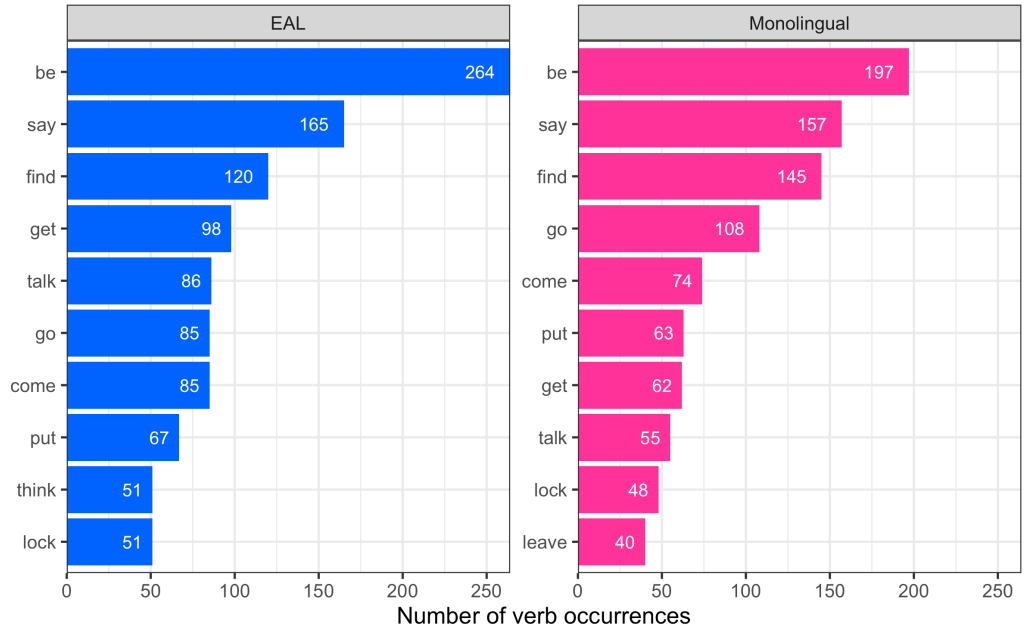


Figure S2b. ***Comparison between EAL and Monolingual groups on 10 most frequent verbs in Year 3.***

Table S3. ***Frequency of clause use by type for EAL and Monolingual groups in Years 1 and 3.***

Clause code	Clause type	EAL		MONO	
		Year 1	Year 3	Year 1	Year 3
m	main	0.66	0.62	0.65	0.56
cr	reported speech	0.07	0.08	0.07	0.08
m+	main with elided subject	0.06	0.07	0.04	0.1
cn	non-finite complement	0.05	0.06	0.04	0.07
ca	causal adverbial	0.04	0.04	0.05	0.05
vp	verb phrase	0.03	0.01	0.02	0.02
a	adverbial	0.02	0.02	0.02	0.03
cf	finite complement	0.02	0.05	0.03	0.02
x	no-verb phrase	0.02	0.01	0.04	0.02
i	imperative	0.01	0.01	0	0.01
n	non-finite, non-complement	0.01	0.01	0.01	0.02
r	relative	0.01	0.02	0.01	0.02
cc	comment clause	0	0	0	0
n+	causal non-finite, non-complement	0	0	0.01	0.01

License

Parents' hyper-pitch and low vowel category variability in infant-directed speech are associated with 18-month-old toddlers' expressive vocabulary

Audun Rosslund
University of Oslo, Norway

Julien Mayor
University of Oslo, Norway

Gabriella Óturai
UiT The Arctic University of Norway, Norway

Natalia Kartushina
University of Oslo, Norway

Abstract: The present study examines the acoustic properties of infant-directed speech (IDS) as compared to adult-directed speech (ADS) in Norwegian parents of 18-month-old toddlers, and whether these properties relate to toddlers' expressive vocabulary size. Twenty-one parent-toddler dyads from Tromsø, Northern Norway participated in the study. Parents (16 mothers, 5 fathers), speaking a Northern Norwegian dialect, were recorded in the lab reading a storybook to their toddler (IDS register), and to an experimenter (ADS register). The storybook was designed for the purpose of the study, ensuring identical linguistic contexts across speakers and registers, and multiple representations of each of the nine Norwegian long vowels. We examined both traditionally reported measures of IDS: pitch, pitch range, vowel duration and vowel space expansion, but also novel measures: vowel category variability and vowel category distinctiveness. Our results showed that Norwegian IDS, as compared to ADS, had similar characteristics as in previously reported languages: higher pitch, wider pitch range, longer vowel duration, and expanded vowel space area; in addition, it had more variable vowel categories. Further, parents' hyper-pitch, that is, the within-parent increase in pitch in IDS as compared to ADS, and lower vowel category variability in IDS itself, were related to toddlers' vocabulary. Our results point towards potentially facilitating roles of increase in parents' pitch when talking to their toddlers and of consistency in vowel production in early word learning.

Keywords: infant-directed speech; vocabulary; vowels; language acquisition; Norwegian

Corresponding author(s): Audun Rosslund, Center for Multilingualism in Society across the Lifespan, University of Oslo, Forskningsveien 3A, 0373, Oslo, Norway. Email: audun.rosslund@iln.uio.no

ORCID ID(s): https://orcid.org/0000-0002-2646-8053; https://orcid.org/0000-0001-9827-5421; https://orcid.org/0000-0002-6526-3392; https://orcid.org/0000-0003-4650-5832

Citation: Rosslund, A., Mayor, J., Óturai, G., & Kartushina, N. (2022). Parents' hyper-pitch and low vowel category variability in infant-directed speech are associated with 18-month-old toddlers' expressive vocabulary. *Language Development Research*, *2*(1), 223—267. https://doi.org/10.34842/2022.0547

Introduction

When talking to infants and young children, adults fine-tune their speech by slowing it down, heightening their pitch, increasing their pitch range and extending their corner vowels (Fernald, 1989; Kuhl et al., 1997). This speech register, known as infant-directed speech (IDS), functions as a 'perceptual hook' and is suggested to aid infants in the task of language acquisition (Cristia, 2013; Golinkoff et al., 2015). Infants prefer listening to IDS over adult-directed speech (ADS) already two days after birth (Cooper & Aslin, 1990), and this preference increases with language exposure, that is, having stronger effects in older infants, and in infants' native over non-native language (The ManyBabies Consortium, 2020), a preference also correlating with relative language exposure in bilingual infants (Byers-Heinlein et al., 2021). However, there are some inconsistencies in the IDS research, in particular with respect to (1) which properties of IDS may facilitate early language development, (2) whether IDS speech is clearer as compared to ADS, (3) the generalisability of the results to different socio-linguistic contexts, and (4) the methods used to record and analyse IDS. Next, we detail each of these points and describe how they are addressed in the current study.

Both experimental and descriptive studies have reported evidence suggesting that IDS may facilitate language development. Experimental studies have shown that stimuli (words and sentences) that imitate prototypical IDS characteristics facilitate word segmentation (Thiessen et al., 2005), word comprehension (Song et al., 2010) and immediate word learning (Graf Estes & Hurley, 2013; Ma et al., 2011). Analogously, descriptive studies linking properties of parents' IDS to children's language outcomes have found positive correlations between vowel space expansion (larger triangular area between the three corner vowels /i/, /ɑ/, /u/ in IDS as compared to ADS) and expressive vocabulary size (Hartman et al., 2017; Kalashnikova & Burnham, 2018), consonant discrimination (García-Sierra et al., 2021; Kalashnikova & Carreiras, 2021; Liu et al., 2003) and complexity of child vocalizations (Marklund et al., 2021), and between pitch range and expressive vocabulary size (Porritt et al., 2014). Larger vowel space expansion has been hypothesised to increase the clarity of speech, thus making sound categories and words (e.g., *bed vs. bad*) easier to distinguish for language learners. This relationship has originally been observed in adult research on speech perception, when vowel space expansion, together with other phonetic features, were found to lead to better speech intelligibility (see e.g., Garnier et al., 2018), hence clear perceived articulation of speech sounds. Yet, increased vowel space expansion *per* se does not necessarily lead to more intelligible speech (for IDS, see Cristia & Seidl, 2014; Miyazawa et al., 2017). Acoustic analyses of parental recordings revealed increased within-category variability in IDS, which might reduce speech clarity (Cristia & Seidl, 2014; Martin et al., 2015; McMurray et al., 2013; Miyazawa et al., 2017). For example, Japanese mothers of 18–20-month-old toddlers extended their first and second formants when talking to their child, as compared to ADS; yet the increased vowel space area did not lead to more distinct categories due to increased variability in vowel

tokens (Miyazawa et al., 2017). Thus, it remains unclear whether the relationship between vowel space expansion in IDS and infants' language outcomes (e.g., Kalashnikova & Burnham, 2018) is attributed to (intentionally) clearer speech provided to the child by the parent, or to a different mediating factor or their combination, such as higher pitch and increased pitch variability, smiling and affect (Benders, 2013), or attempts to appear smaller and less intimidating to the child (Kalashnikova et al., 2017), all of which might potentially lead to vowel space expansion.[1]

Another central question is whether the acoustic properties of IDS – and the potential boosting effect of certain IDS properties in language acquisition – are similar across different socio-linguistic contexts, that is, cultures with varying parenting behaviours, and languages and dialects with varying linguistic structures. As detailed below, this is likely not the case (Saint-Georges et al., 2013, and see e.g., Casillas et al., 2020; Cristia et al., 2022 for descriptions of cultures with infrequent child-directed vocalisations). The majority of studies on IDS have been conducted with American English parents (for the overall prevalence of English in child language studies, see Kidd & Garcia, 2022), who have been described as having more extreme IDS properties than parents in other languages might display (Fernald et al., 1989), questioning the generalizability of the results. While higher and more variable pitch might be the two most robust characteristics of IDS present across most cultures and languages (Broesch & Bryant, 2015; Farran et al., 2016; McClay et al., 2021; Narayan & McDermott, 2016; but see also Han et al., 2020, 2021), vowel space expansion, on the other hand, has not been reported consistently across languages. For instance, increased vowel space expansion in IDS *vs.* ADS has not been found in Dutch (Benders, 2013), German (Audibert & Falk, 2018), Cantonese (Xu Rattanasone et al., 2013), Lenakel and Southwest Tanna (McClay et al., 2021), and reported inconsistently for Norwegian (Englund & Behne, 2006; Kartushina et al., 2021). Further, experimental studies have found that neither British (Floccia et al., 2016) nor German (Schreiner & Mani, 2017) infants segment speech stimuli recorded in natural IDS register in their respective languages, unless these were prosodically exaggerated over and beyond what would be considered 'natural' British and German IDS. Overall, these findings paint the picture that IDS and its potential effect on language development are not uniform, and call for studies of IDS across a wider range of languages and dialectal variations.

[1] We deliberately avoid the term 'hyperarticulation' throughout this manuscript. Although vowel space expansion is, originally, the acoustic proxy for 'hyperarticulation', it is, yet, a component of clear speech; in infant development research, the term is often used interchangeably with clear speech *per se*, not acknowledging potential underlying variability in sound production that may make speech less clear (cf references in the text).

A final concern is the varying procedures used to elicit IDS and to measure its acoustic properties. For example, IDS (and ADS) have been recorded in both home (Narayan & McDermott, 2016) and lab-environments (Benders, 2013), during unstructured (Englund & Behne, 2006) or semi-structured interactions (Kalashnikova & Burnham, 2018), elicited through a picture-description task (Weirich & Simpson, 2019) or a storybook reading (Burnham et al., 2015; McMurray et al., 2013). These differences in the recording contexts can influence the acoustic properties of speech (e.g., Burnham et al., 2015; Miyazawa et al., 2017; Tamis-LeMonda et al., 2017); thus, researchers should weigh the pros and cons of each procedure. In addition, researchers can examine the acoustic properties of parental speech when addressed to their child, the IDS *per se* (e.g., Hartman et al., 2017; Liu et al., 2003; Porritt et al., 2014) or the within-parent difference between the acoustic measures of IDS as compared to ADS, meaning that parents function as their own baseline (e.g., Kalashnikova & Burnham, 2018; Kalashnikova & Carreiras, 2021). Given that these two lines of research in fact capture two complementary constructs of parents' speech – the acoustic features of IDS, and the acoustic difference between the two registers (or the perceived 'adaptation', whether parents modulate it, consciously or not) – there is a need for integrative studies that combine both approaches and examine their respective contribution to the child's early language development.

Hence, the aims of this study were three-fold. First, we sought to assess IDS in comparison to ADS in Norwegian parents speaking a Northern Norwegian dialect. To elicit IDS, we designed a child-friendly storybook[2] (see Methods for details) that enabled us to collect 10 vowel tokens, varying in surrounding consonantal context (5 types), for each of the 9 Norwegian long vowels, providing a more comprehensive analysis of vowels addressed to the child, as compared to describing the three 'corner' vowels in previous research (as also criticised by e.g., Englund, 2018). Parents read this book to their 18-month-old toddler (IDS), as well as to another adult (ADS). This procedure ensured that elicited speech was sampled from identical linguistic contexts across the two registers and speakers (Steinlen & Bohn, 1999; Wang et al., 2015), providing better generalizability across the registers. We examined the acoustic measures of speech that are traditionally reported: that is, pitch, pitch range, vowel duration and vowel space area (Fernald, 1989; Kuhl et al., 1997; Wang et al., 2015), but also novel measures of vowel category variability and vowel category distinctiveness, providing novel proxies/indices for the clarity of speech, as an increased vowel space might also

[2] Note that, for the sake of simplicity, we refer to the storybook-elicited speech read to a child and to an adult as IDS and ADS, respectively.

contain more variability within each vowel category and, hence, lead to less distinct vowel categories (see e.g., Cristia & Seidl, 2014; McMurray et al., 2013). Second, we aimed to evaluate whether the within-parent differences – or adaptation – between IDS and ADS, if any, predicted the expressive vocabulary size of their 18-month-old toddlers (similarly to e.g., Kalashnikova & Burnham, 2018). Finally, we sought to assess whether any of the acoustic measures examined in the current study for IDS, *not* the difference between registers, or adaptation, predicted toddlers' expressive vocabulary (similarly to e.g., Hartman et al., 2017). It is noteworthy that Norwegian language uses vowel formants, vowel length and pitch accent as cues to mark lexical meaning. In addition, Norway is characterised by its dialect diversity, with differences in lexicons, phonemic realisation, and pitch accent patterns across dialects (Mæhlum & Røyneland, 2012). Given that the current knowledge about IDS in Norwegian comes from speakers of the Central Norwegian dialect (Englund, 2018; Englund & Behne, 2005, 2006), the current study (with speakers of the Northern Norwegian dialect) may also highlight potential diversity of IDS in a more fine-grained manner, that is, within-language, but across-dialect.[3]

For our first aim, and in line with previous studies, we expected, as per pre-registration (https://osf.io/7st6w/), that when addressing speech to their child (IDS), in comparison to an adult (ADS), Norwegian parents will produce: higher pitch, wider pitch range and increased vowel duration. With respect to the vowel space area, Englund & Behne (2006) found a decrease in Norwegian parents' IDS addressed to 1–6-month-old infants, whereas Kartushina and colleagues (2021) found an increase in Norwegian parents' IDS addressed to 8-month-old infants. These differences in vowel space can be due to either children's ages (0–6-month-olds *vs.* 8-month-olds), differences in dialects (Central *vs.* Eastern Norwegian), or methods to compute vowel space (using /ɑ:/ *vs.* /æ:/ as the extreme/corner open vowel in Englund & Behne, 2006 and Kartushina et al., 2021, respectively), or a combination of these factors. Given that the current study examined parents speaking a Northern Norwegian dialect directed to older toddlers, and measured vowel space using the /æ:/ vowel as the most extreme open vowel in Norwegian, we predicted, in line with Kartushina and colleagues (2021), vowel space expansion in IDS, as compared to ADS. Finally, in line with recent results in Norwegian (Kartushina et al., 2021), English (Cristia & Seidl, 2014; McMurray et al.,

[3] We note that distinguishing dialects from languages is not necessarily linguistically meaningful, as this distinction is primarily linked to political and cultural factors (yet, for a recent attempt, see Wichmann, 2020).

2013) and Japanese (Martin et al., 2015; Miyazawa et al., 2014), we expected vowel categories to be less compact and less distinct in IDS, as compared to ADS.

For our second aim, to evaluate whether the within-parent differences – or adaptation – between IDS and ADS, if any, predict the vocabulary of their toddlers, in line with previous research, we expected that increases in pitch, pitch range and vowel duration would be positively related to toddlers' expressive vocabulary. Given that pitch accent and vowel duration are lexically meaningful cues in Norwegian (they are used to distinguish words, as, for example in *tak* [roof] *vs. takk* [thanks] or *bønder* [farmer] *vs. bønner* [beans]), we expected that toddlers would benefit from input that emphasises these cues in IDS, especially since, at 18 months of age, their expressive vocabulary is rapidly increasing. In addition, we expected a positive relationship between vowel space expansion and toddlers' expressive vocabulary (as found in Hartman et al., 2017; Kalashnikova & Burnham, 2018). Finally, as we expected increased within-vowel category variability and less between-vowel distinctiveness in IDS, as compared to ADS (e.g., Cristia & Seidl, 2014; McMurray et al., 2013), we anticipated that parents who produce less variable and/or more distinct vowel categories would, by means of facilitating speech sound discrimination and representations, boost their child's word learning. Hence, we expected a negative relationship between vowel category variability and toddlers' expressive vocabulary, but a positive relationship between vowel category distinctiveness and toddlers' expressive vocabulary. To summarise, we hypothesise that the 'ideal' IDS adaptation benefiting early word learning contains exaggerated (a) pitch and pitch range, (b) vowel duration, and (c) vowel space, and (d) precise vowel tokens with (e) little variability within each category.

Last, and for our third aim, we assessed whether any of the acoustic measures examined in the current study for parents' IDS itself, *not* the difference between the registers, predicted toddlers' vocabulary, and we expected that the same acoustic features as those that were emphasised in IDS when compared to ADS (within-parent differences between the registers), would be associated with toddlers' expressive vocabulary. That is, parent-specific pitch, pitch range, vowel duration, vowel space area and vowel category distinctiveness in IDS would be positively related to toddlers' expressive vocabulary, while vowel category variability would be negatively related to toddlers' expressive vocabulary.

Method

Participants

Twenty-one parent-toddler dyads from the city of Tromsø (Northern Norway) participated in the current study. Two additional dyads were recruited, but excluded from the analysis, due to missing audio files (n = 1) and less than 75% exposure to Norwegian (n = 1). For the final sample, all parents (16 mothers, 5 fathers) were native

speakers of Norwegian, raised in Northern Norway and spoke the Northern Norwegian dialect. All parents cohabited with their toddlers and the toddlers' other parent, and reported to provide at least 50% of speech input to their toddler as compared to the other parent. Toddlers (9 girls, 12 boys, *M* age = 17.9 months, *SD* = 0.43) were exposed, on average, to 97.5% of Norwegian (*SD* = 7.49) and none had reported any visual or auditive impairments.[4] Socioeconomic status (SES), reported as mother's highest education level, ranged from 1 (secondary school) to 5 (doctoral degree), with the median being 3 (bachelor's degree).

Data collection took place in the BabyLab at the Department of Psychology, University of Tromsø. After receiving invitations through advertisement on social media, at the university, local library or health station, parents who agreed to participate with their child in the study signed an informed consent form, and within the five days after their visit to the lab, answered a web questionnaire that included general demographic questions and questions about their toddlers' linguistic environment. The online questionnaire included the Norwegian adaptation of the MacArthur-Bates Communicative Development Inventories (CDI) –Words and Sentences form (Simonsen et al., 2014). Individual raw CDI scores (the number of words that parents reported their child to produce) were converted to daily percentiles using the normative Norwegian data from Wordbank (Frank et al., 2017; for the conversion procedure, see Kartushina al., 2022); the mean score was 37.6 (*SD* = 29.3, range = 1–93). The current study was conducted according to the guidelines laid in the Declaration of Helsinki, with written informed consent obtained from a parent or a guardian for a child before any assessment or data collection. The study has been approved by the Norwegian Centre for Research Data (NSD, ref. 56312), and the local ethical committee at the Department of Psychology, University of Oslo. The pre-registration, data, stimuli and analysis script for the study are openly available at the Open Science Framework (OSF) project's page (https://osf.io/7st6w/).

[4] Two of the toddlers were reported to be born 'too early'. The exclusion criteria for toddlers was to be born before 37 weeks of gestation (i.e., premature according to medical convention). However, poor wording of this specific question in our questionnaire made parents' responses ambiguous. The wording of the question was open, not specific to the number of weeks and did not include the term 'premature'. Thus, we were not able to know whether these two toddlers were in fact premature or simply born any time (e.g., one or two days) before the expected due date. Comparing these two toddlers to the rest of the sample on the key measures did not reveal any differences (see Appendix 2). We, therefore, included them for the analyses.

Procedure and Stimuli

Upon the arrival to the BabyLab, parents and their toddlers were familiarised with the lab environment and experimenters and received information about the course of their visit. Seven of the toddlers took part in an unrelated experiment on motor imitation prior to the recordings. Parents were not aware of the specific purpose of the study, or which parts of their recorded speech were of interest to the researchers, until after they had completed the recording sessions.

The IDS and ADS recordings took place either in the waiting area of the BabyLab, or in an adjacent child-friendly room. Both IDS and ADS were elicited from the parent through reading a child-friendly storybook, specifically created for the purpose of the study. The storybook was written in Norwegian Bokmål[5] and consisted of five pages, 39 sentences and 327 words. Each page had a colourful illustration and a short child-friendly narrative (Table 1); the narratives were not connected with each other. The nine long Norwegian vowels (/ɑ:/, /e:/, /i:/, /u:/, /ʉ:/, /y:/, /æ:/, /ø:/, and /ɔ:/) were represented by five unique words repeated twice throughout the storybook, for a total of 90 target vowels. The words were mono- and bisyllabic lexical and function words, most of them reported to be known by a large proportion of toddlers at this age (Simonsen et al., 2014). Words were counterbalanced in terms of their position within a sentence, so that each target vowel was present in at least one start-, mid- and end-sentence word. The target vowel was in a stressed position within the word, and, for the bisyllabic words, with the two exceptions, the target vowel was always placed in the first syllable. See Appendix 1 for an overview of target vowels within words.

During the IDS recording, the parent read the storybook to their toddler either sitting on their lap or next to them. Parents were instructed to read and interact with their child as they would typically do when reading a book at home. Parents did not receive any instructions with respect to the dialect to use (recall the book was written in Norwegian Bokmål, which is close to the Eastern, Oslo-area, dialect); all parents chose to read in their Northern Norwegian dialect, that is, adapting the grammatical gender, the phonemic realisation, and the intonation patterns to this dialect. During the ADS recording, parents read the same storybook to the experimenter (a native speaker of Norwegian), with no further instructions but to read the book naturally as if reading to an adult. Again, parents chose to read in their Northern Norwegian dialect. During

[5] Dialects are not used in written text; hence this is one of two official, dialect-neutral, written forms of Norwegian.

the ADS recording, a second experimenter cared for the toddler outside of the parents' field of vision. Due to limited resources, the second experimenter was not available for three parent-toddler dyads. The order of the recordings was counterbalanced; half of the parents started with the IDS, and the other half started with the ADS. All sessions were recorded with an Olympus DS-3000 handheld voice recorder in 16-bit/44.1 kHz. After the recordings, toddlers received a small toy or a book as a token of appreciation.

Table 1. ***Example of text from one page in the storybook (words with target vowels in bold, IPA transcripts in brackets)***

Original	English translation
Mamma-sjiraffen **skjærer** [ʂæːrer] en **skive** [ʂiːvə] av **brødet** [brøːə]. Den lille sjiraffen ligger på **magen** [mɑːgən], med den ene **foten** [fuːtən] i været. Han vil heller ha **kake** [kɑːkə] og **banan** [bɑnɑːn]. Mamma-sjiraffen **skjærer** [ʂæːrer] enda en **skive** [ʂiːvə] av **brødet** [brøːə], og legger fram en **skje** [ʂeː] til grøten. "Vi kan **spise** [spiːse] **kake** [kɑːkə] og **banan** [bɑnɑːn] etterpå", sier mamma-sjiraffen. "**Bra**! [brɑː]", sier den lille sjiraffen.	Mommy-giraffe **cuts** a **slice** of **bread**. The little giraffe is lying on his belly, with one **foot** in the air. He would rather have **cake** and **banana**. Mommy-giraffe **cuts** another **slice** of **bread**, and lays out a **spoon** for the porridge. "We can **eat cake** and **banana** later", says Mommy-giraffe. "**Great**!", says the little giraffe.

Data Processing and Acoustic Measures

Three trained native speakers of Norwegian listened to the audio recordings in Praat (Boersma & Weenink, 2020) and marked the target speech segments. First, they segmented parents' speech and marked the onset and the offset of the phrases, necessary for the pitch analyses. A phrase was defined as a portion of continuous speech with intact pitch tracks, without interruptions (e.g., interference from the child), enclosed by approximately 500 ms of silence, typically a pause where the parent drew breath. In other words, the length and the content of a phrase varied across segments and could include short utterances as well as full sentences. In total, we identified 923 phrases in IDS and 818 phrases in ADS. A customised Praat script (Hirst, 2012) automatically extracted the duration and the minimum, maximum, and mean pitch (F0) in Hz for each phrase. 133 phrases (7.6%) were manually corrected due to errors in the octave jumps (i.e., pitch tracks printed one octave higher than intended). As pitch perception follows a logarithmic scale, all Hz values were converted to semitones using the following formula $semitones = 12*\log^2(F0/constant)$, as in Kalashnikova & Burnham (2018), with 10 as a constant (i.e., semitones-above-10-hertz). Pitch range was

computed as the difference between the minimum and the maximum pitch value (in semitones) within each phrase.

Second, we identified and manually annotated the target vowels. Only audible target vowels, with a minimum length of 30 ms, with no noise and with visually trackable first (F1) and second (F2) formants were segmented. We followed the same vowel onset and offset boundary definition as in Cristia & Seidl (2014). In total, we identified 1577 vowels in IDS and 1527 vowels in ADS. A customised Praat script (Hirst, 2012) was run to collect vowel duration (in ms) and the mean F0, F1 and F2 (in Hz), with the pre-specified formant ceiling values at 5500 Hz for mothers and 5000 Hz for fathers. 297 vowel segments (9.6%) were manually corrected due to errors in the formant estimates (typically identifying F1 as F2, or F3 as F2, which could be due to high F0, see Monsen & Engebretson, 1983). See Figure 1 and Table 2 for an overview of all vowel segments. Computations of the different vowel-based measures are explained below.

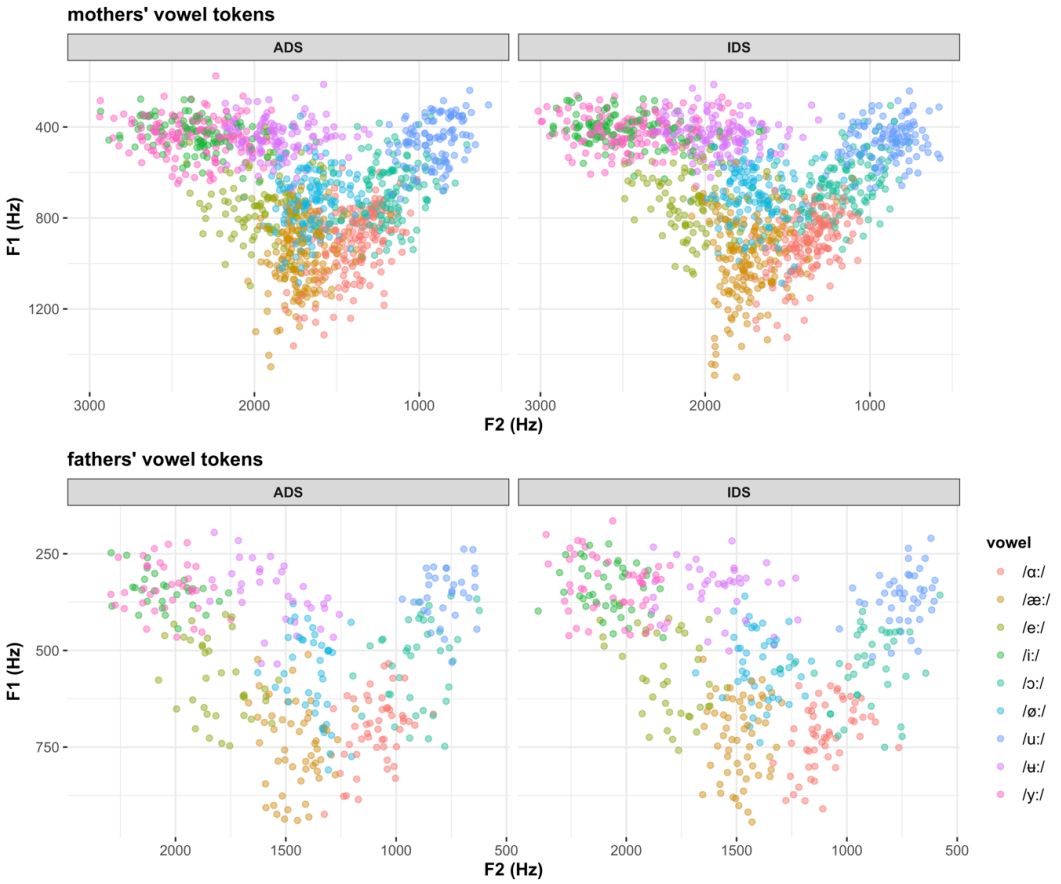


Figure 1. ***Mother's and father's vowel tokens in F1-F2 space by register***

Table 2. ***Number of tokens, mean duration (ms) and mean formant frequencies (Hz) for each target vowel across IDS and ADS registers for mothers and fathers, with standard deviations in parentheses***

	ADS								**IDS**							
	mothers				fathers				mothers				fathers			
	n	duration	F1	F2	n	duration	F1	F2	n	duration	F1	F2	n	duration	F1	F2
/iː/	107	96.0	434	2350	29	123	529	896	118	114	417	2460	40	116	338	2080
		(30.1)	(71.9)	(244)		(43.7)	(112)	(132)		(58.9)	(69.9)	(272)		(48.2)	(63.9)	(129)
/yː/	132	102	442	2390	36	120	340	2040	128	128	415	2470	48	106	331	2070
		(29.9)	(88.7)	(227)		(35.8)	(64.8)	(131)		(39.3)	(75.5)	(265)		(31.0)	(74.9)	(159)
/eː/	97	109	747	1960	36	99.5	576	1800	92	120	734	2080	38	107	564	1820
		(45.3)	(135)	(192)		(35.8)	(101)	(140)		(50.2)	(139)	(191)		(35.2)	(108)	(144)
/øː/	137	111	728	1660	40	117	566	1380	130	121	701	1690	46	113	539	1390
		(36.0)	(130)	(140)		(27.1)	(113)	(89.7)		(40.5)	(127)	(159)		(21.1)	(85.3)	(104)
/æː/	183	106	951	1710	53	113	747	1450	176	119	974	1730	58	115	725	1480
		(37.8)	(147)	(135)		(33.3)	(108)	(109)		(50.3)	(182)	(167)		(44.7)	(102)	(91.3)
/ʉː/	123	108	448	1860	34	102	358	1550	131	127	432	1920	41	103	349	1580
		(44.4)	(80.4)	(196)		(30.7)	(82.8)	(164)		(53.1)	(78.1)	(209)		(25.8)	(70.5)	(181)
/uː/	110	133	461	909	32	119	353	775	113	169	462	894	42	143	364	747
		(55.7)	(105)	(133)		(30.4)	(67.6)	(89.7)		(86.5)	(89.5)	(152)		(60.2)	(68.9)	(102)
/ɔː/	128	112	683	1190	39	122	529	896	126	140	656	1180	39	125	542	902
		(38.0)	(130)	(169)		(43.7)	(112)	(132)		(60.3)	(125)	(201)		(36.7)	(98.7)	(149)
/ɑː/	162	138	911	1400	49	135	707	1090	157	171	907	1380	54	154	717	1090
		(57.8)	(153)	(165)		(38.8)	(90.2)	(104)		(89.9)	(131)	(151)		(67.1)	(86.5)	(113)
Mean	131	113	672	1700	38.7	117	529	1420	130	135	654	1740	45.1	121	510	1460
	(27.3)	(44.6)	(237)	(481)	(7.8)	(36.7)	(181)	(435)	(24.4)	(64.9)	(246)	(533)	(7.01)	(47.2)	(178)	(462)

Vowel Space Area

For the vowel space area (VSA), we measured the overall size of the F1-F2 vowel space (in Hz^2) with the *phonR* package (McCloy, 2016), using the average F1 and F2 (in Hz) for each vowel category and the following formula (exemplified with three vowels, where 'ABS' is the absolute value): $ABS\ ½ \times [(F1/vowel_1/ \times (F2/vowel_2/ - F2/vowel_3/) + F1/vowel_2/ \times (F2/vowel_3/ - F2/vowel_1/) + F1/vowel_3/ \times (F2/vowel_1/ - F2/vowel_2/)]$ and so forth, previously used in IDS research (Kalashnikova & Burnham, 2018; Kuhl et al., 1997; Liu et al., 2003). For each register and each parent, we computed three different vowel space area (VSA) measures: one using the corner vowels /i:/, /**ɑ**:/, /u:/ ("VSA_a"), in line with previous research in IDS, including Norwegian (Englund, 2018); one using the corner vowels /i:/, /**æ**:/, /u:/ ("VSA_æ"), as, based on earlier findings in Norwegian (Kartushina et al., 2021) and also confirmed by our data, /æ:/ is the most extreme Norwegian open vowel in the F1-F2 space (see Figure 1). In addition, we computed a measure of vowel space area including all border vowels; /ɑ:/, /e:/, /i:/, /u:/, /ʉ:/, /æ:/, /ɔ:/ ("VSA_full"), as this would measure most accurately the total vowel area, as the actual vowel space may not necessarily be accurately represented by a triangle.

Vowel Category Variability

The vowel category variability score is an index of the within-category precision in vowel production.[6] The variability of each vowel category in the F1-F2 vowel space (as also used by Hartman and colleagues, 2017) was measured by fitting F1 and F2 (Hz) of all vowel tokens, exemplifying the category, to a customised MatLab script (Kartushina & Frauenfelder, 2014), which calculated the area of an ellipse (Hz^2) for each vowel category, participant, and register, with the following formula: $ellipse\_area = F1 \times F2 \times \pi$, where σF1 is 1 standard deviation of the mean of F1, and σF2 is 1 standard deviation of the mean of F2. Since the distribution of the productions in F1/F2 space was assumed to be elliptical, we estimated the angles of the major and minor axes of an ellipse centered on the mean of the productions (in order to determine the orientation of the axes). Therefore, a low vowel category variability score indicated more compact vowel categories, whereas a high vowel category variability score indicated looser vowel categories.

[6] Note that in the pre-registration, we referred to this measure as 'vowel category compactness'.

Vowel Category Distinctiveness

For vowel category distinctiveness, we measured how distinct participants' vowel categories were from each other in the F1-F2 vowel space. Thus, while vowel category variability indicates the precision of vowel production within each category, vowel category distinctiveness indicates the discriminability of the categories, i.e., the degree of overlap, taking into account their distribution within the full vowel space. Vowel category distinctiveness was computed as the between-vowel category Sum of Squares (the squared distances of category cluster centroids from the overall vowel space centroid) divided by the total Sum of Squares (squared distances of individual vowel tokens from the overall vowel space centroid), for each participant and register, for 8 vowel categories (we omitted the category /y/, as it fully overlaps with the Norwegian /i/ in the F1-F2 space, as the distinguishing feature is F3). See Appendices 3A and 3B for a thorough explanation and visual representation of the measure as a function of the amount of overlap between the vowel categories. Thus, vowel category distinctiveness can be thought of as a clustering performance quotient, indexing the proportion of variance in F1 and F2 explained by the vowel category identity, ranging from 0 (cluster/category membership explains no variance) to 1 (cluster/category membership explains all variance). In sum, with these three F1-F2 based measures, computed across vowel categories, we aimed to thoroughly describe the distinguishing features of parents' vowel production in IDS. For further details on the computation of measures, we refer readers to the available code on the OSF project page (https://osf.io/7st6w/).

Results

The results are structured according to the three aims of the current study; 1) to examine whether there were differences in acoustic properties, both traditional (pitch, pitch range, vowel duration and vowel space area) and novel (vowel category variability and vowel category distinctiveness), between IDS and ADS, 2) to assess the role of within-parent differences between the IDS and ADS registers in predicting toddlers' expressive vocabulary, and 3) to assess the role of acoustic properties of IDS in predicting toddlers' expressive vocabulary. All analyses were preregistered and conducted in R (R Core Team, 2020), with libraries and their versions listed in Appendix 4.

Acoustic Properties of IDS and ADS

Between-register differences in the acoustic measures were assessed with a linear mixed-effect model separately for each acoustic measure. The fixed structure was similar for all models and included register, parent gender and their interaction; the random structure included participant, as well as register and vowel category for some models (cf details below). Models were fitted with the *lme4* package (Bates et al.,

2015) and the model assumptions, including normality and homogeneity of residuals, were visually inspected on diagnostics plots derived from the check_model() function from the *performance* package (Lüdecke et al., 2021). Models were analysed with the Anova() function from the *car* package (Fox & Weisberg, 2018) with the p-values obtained from the *lmerTest* package, using Satterthwaite approximation (Kuznetsova et al., 2017). All model results are shown in Table 3, and between-register differences are visualised in Figure 2.

Pitch

As shown in Table 3, there was a significant effect of register and parent gender on pitch. That is, as expected, parents had a higher mean pitch (all reported in semitones) in IDS (M = 54.1, SD = 6.67) than in ADS (M = 51.4, SD = 6.02), Hedges g = 1.28. Further, mothers had overall higher mean pitch (M = 55.8, SD = 3.51) than fathers (M = 43.9, SD = 4.93). The register by parent gender interaction was not significant.

Pitch Range

As shown in Table 3, there was a significant effect of register on pitch range: As expected, parents had a wider pitch range (all reported in semitones) in IDS (M = 14.6, SD = 6.39) than in ADS (M = 13.3, SD = 5.59), Hedges g = 0.44. The main effect of parent gender on pitch range, and the interaction effect of parent gender and register were not significant.

Vowel Duration

As shown in Table 3, there was a significant effect of register on vowel duration. Note that we log-transformed the outcome measure in the linear mixed-effects model, because the initial model violated the assumption of normality of residuals (see pre/post diagnostics plots in Appendix 5A and 5B). That is, as expected, parents produced longer vowels (reported in ms here for ease of interpretation) in IDS (M = 131, SD = 61.1) than in ADS (M = 114, SD = 43), Hedges g = 1.05. However, as can be seen in the follow-up analyses using *lsmeans* (Lenth, 2016), the main effect is due to the mothers prolonging their vowels to a greater degree in IDS (M = 135, SD = 64.9) as compared to ADS (M = 113, SD = 44.6, t(16.4) = -5.7, p = < .001), whereas fathers' vowel duration did not differ significantly between the registers (IDS: M = 121, SD = 47.2, ADS: M = 117, SD = 36.7, t(19.6) = -0.3, p = .766).

Vowel Space Area

As shown in Table 3, there was a significant effect of register on all three of our vowel space area measures. To facilitate the descriptive statistics, vowel space areas (reported in Hz^2) were divided by 1000, hence, kHz^2. As expected, parents expanded their

vowel space area in IDS (VSA_a: *M* = 339, *SD* = 99.7; VSA_æ: *M* = 379, *SD* = 124; VSA_full: *M* = 441, *SD* = 113) as compared to ADS (VSA_a: *M* = 303, *SD* = 104; VSA_æ: *M* = 335, *SD* = 106; VSA_full: *M* = 389, *SD* = 120), Hedges *g* = 0.58; 0.55; 0.54, for VSA_a, VSA_æ and VSA_full, respectively. Further, for all vowel space area measures, mothers had overall larger vowel space areas (VSA_a: *M* = 349, *SD* = 97.7; VSA_æ: *M* = 389, *SD* = 44.4; VSA_full: *M* = 445, *SD* = 112) than fathers (VSA_a: *M* = 232, *SD* = 46.4; VSA_æ: *M* = 253, *SD* = 11.3; VSA_full: *M* = 322, *SD* = 88.4). The register by parent gender interaction was not significant for any measure of vowel space.

Vowel Category Variability

As shown in Table 3, there was a significant effect of register and parent gender on vowel category variability. Note that we log-transformed the outcome measure in the linear mixed-effects model, because the initial model violated the assumption of normality of residuals (see pre/post diagnostics plots in Appendix 6A and 6B). To facilitate the interpretability of the descriptive statistics, we report the non-log transformed vowel category variability in kHz^2. As expected, parents had more variable categories in IDS (*M* = 311, *SD* = 225) than in ADS (*M* = 273, *SD* = 0205), Hedges *g* = 0.44. Further, mothers had overall more variable categories (*M* = 333, *SD* = 228) than fathers (*M* = 161, *SD* = 80.4). The register by parent gender interaction was not significant.

Vowel Category Distinctiveness

As shown in Table 3, parent gender was the only significant effect on vowel category distinctiveness, with mothers having overall less distinct categories (*M* = 0.88, *SD* = 0.04) than fathers (*M* = 0.93, *SD* = 0.02), Hedges *g* = -1.50. Contrary to our expectation, there were no differences between the two registers, and the register by parent gender interaction was not significant.

Table 3. ***Model outputs on acoustic differences between the IDS and ADS registers (n = 21 parent-toddler dyads)***

Model	Parameter	2	*df*	*p*
Pitch ~	Register	40.72	1	<.001***
Register * Gender +	Gender	127.2	1	<.001***
(1 + Register | Participant)[7]	Register * Gender	2.209	1	.137
Pitch_range ~	Register	4.308	1	.038*
Register * Gender +	Gender	1.016	1	.314
(1 + Register | Participant)[7]	Register * Gender	0.121	1	.728
Vowel_duration ~	Register	25.09	1	<.001***
Register * Gender +	Gender	0.159	1	.690
(1 + Register | Participant) + (1 + Register | Vowel)[8]	Register * Gender	8.020	1	.005**
Vowel_space_a ~	Register	7.559	1	.006**
Register * Gender +	Gender	7.541	1	.006**
(1 | Participant)	Register * Gender	0.638	1	.424
Vowel_space_æ ~	Register	7.351	1	.007**
Register * Gender +	Gender	8.077	1	.004**
(1 | Participant)	Register * Gender	2.389	1	.122
Vowel_space_full ~	Register	6.656	1	.010*
Register * Gender +	Gender	6.298	1	.012*
(1 | Participant)	Register * Gender	0.982	1	.322
Vowel_category_variability ~	Register	8.891	1	.003**
Register * Gender +	Gender	7.700	1	.006**
(1 | Participant) + (1 + Register | Vowel)[8]	Register * Gender	0.203	1	.652
Vowel_category_distinctiveness ~	Register	0.001	1	.977
Register * Gender +	Gender	9.683	1	.002**
(1 | Participant)	Register * Gender	0.067	1	.796

*p < .05, **p < .01, ***p < .001

[7] Note that these models deviate from that specified in the pre-registration, where we included a random structure of the segmented phrase in which we extracted the pitch tracks. Given that the number of phrases and their content varied across registers, it was impossible to have similar segment structures.

[8] Recall that the outcome variable was log-transformed.

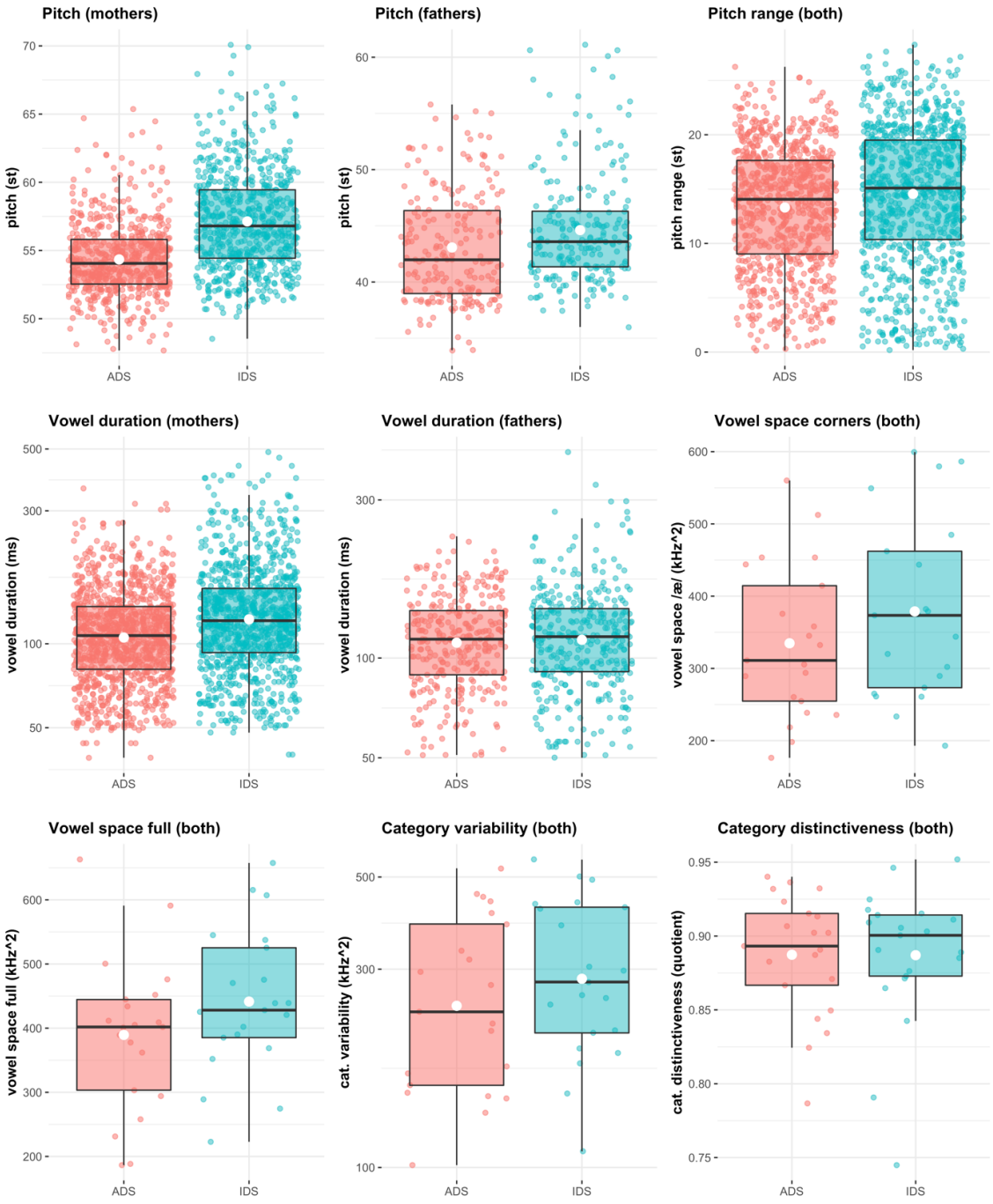


Figure 2. ***Boxplots of acoustic measures in IDS and ADS. Note that the white dots represent the mean. Pitch and vowel duration are visualised separately for parent gender. For vowel duration and category variability, y-axis ticks indicate the scale in the original units, but data is plotted with log-transformed units as this was used in our models. Pitch and pitch range are in semitones, vowel duration in milliseconds, vowel spaces and category variability in kHz² and category distinctiveness in quotients.***

Within-Parent Differences Between IDS and ADS and Toddlers' Expressive Vocabulary

To assess whether the differences parents may have in IDS as compared to ADS predicted toddlers' vocabulary, we computed, first, the ratio between the registers for all the examined acoustic measures, by dividing, for each parent, the average IDS measure by the respective average ADS measure. One exception to this was the vowel space measures – as there was only one measure per register, we did not have to compute the average. A score above 1 indicated a higher value of a specific acoustic measure in IDS, that is, a *hyper*-feature in IDS, and a score below 1 indicated a higher value of a specific acoustic measure in ADS, that is, a *hypo*-feature in IDS. Next, we z-transformed these ratios for each acoustic measure, to facilitate model convergence. Finally, we fitted a beta-regression model using the *betareg* package (Cribari-Neto & Zeileis, 2010), with the outcome measure toddlers' CDI percentiles divided by 100, as required for the beta distributions. The model parameters were:[9]

> CDI percentile ~ Pitch diff_z + Pitch range diff_z + Vowel duration diff_z + Vowel space_æ diff_z + Vowel space_full diff_z + Vowel category variability diff_z

As can be seen in the model output (produced by the *summary* function on the model) reported in Table 4, parents' pitch difference significantly predicted toddlers' vocabulary in percentiles, whereas the other acoustic measures did not. As visualised in Figure 3, parents' hyper-pitch, i.e., an increase in IDS as compared to ADS, was positively related to vocabulary, that is, CDI percentiles increased by 0.71 when pitch difference increased by one standard deviation of the sample mean with all other factors kept at an average. To examine if such an increase in pitch was a deliberate choice parents made, we computed, in an exploratorily analysis, a correlation between parents' hyper-pitch and a mean score of four items retrieved from our background

[9] Given that some of our acoustic measures were highly correlated, such as the two measures of vowel space (using corner vowels versus using the full vowel space), and vowel category variability and vowel category distinctiveness (see Appendix 7), we used the variance inflation factor (VIF) to estimate multicollinearity between the predictors. We took a conservative approach and kept predictors within the VIF < 2.5 (e.g., Zuur et al., 2010). Fitting the pre-registered model resulted in high VIFs for the vowel category variability (VIF = 3.01) and vowel category distinctiveness (VIF = 3.29), and so we excluded the latter, given that we did not find any differences between parents' category distinctiveness across registers.

questionnaire that examined parental attitudes towards early language development, developed in Frank and Hembacher (2020)[10], finding no significant relationship, $r_s(19) = .25$, $p = .275$, suggesting that parents' variability in hyper-pitch in IDS was not related to their differences in beliefs that parents need to provide salient linguistic input in an infant-friendly manner to their child.

Table 4. ***Beta-regression results for vocabulary by parents' difference in IDS vs. ADS (n = 21 parent toddler dyads)***

Parameter	*estimate*	*SE*	z	*p*
Intercept	-0.500	0.203	-2.464	.014*
Pitch diff_z	0.705	0.235	3.008	.003**
Pitch range diff_z	0.313	0.237	1.320	.187
Vowel duration diff_z	0.124	0.235	0.529	.597
Vowel space_æ diff_z	0.564	0.317	1.780	.075
Vowel space_full diff_z	-0.325	0.303	-1.075	.283
Vowel category variability diff_z	0.406	0.230	1.766	.077

*$p < .05$, **$p < .01$

Acoustic Properties of Parents' IDS and Toddlers' Expressive Vocabulary

Finally, to assess whether the acoustic properties of parental input in IDS predicted toddlers' vocabulary, independently of any differences between the IDS and ADS registers, we z-transformed mean values on all our acoustic measures in IDS, separately for mothers and fathers. Given that there are physical differences between males and females impacting the acoustics of speech, this was necessary so that, for example, lower pitch and smaller vowel spaces in fathers would not cloud any results. This approach is a deviation from our pre-registered pipeline, where we suggested, 1) to run the model with mothers only, 2) to transform F1 and F2 from Hz to Bark to normalise, then recompute vowel-based measures. The latter did not seem to adjust for between-gender differences as well as predicted. Hence, we chose to instead standardise

[10] The items were the following statements (responses indicating level of agreement on a 0-6 scale): 'Parents can help babies learn language by talking to them' / 'When speaking to a young child, I often speak slower and more clearly' / 'Reading books to children is not useful until they have learned to speak' (reverse coded) / 'When speaking to a young child, I often use a different voice with a more lively tone.'

measures within each gender group. As before, we fitted and analysed a beta-regression model with toddlers' CDI percentiles divided by 100 as our outcome measure. The model parameters were:[11]

> CDI percentile ~ Pitch IDS_z + Pitch range IDS_z + Vowel duration IDS_z + Vowel space_full IDS_z + Vowel category variability IDS_z

The model output can be seen in Table 5. Vowel category variability in IDS significantly predicted toddlers' vocabulary size, whereas the other acoustic measures were not significant. As visualised in Figure 4, parents with more variable vowel categories in IDS had toddlers with lower vocabulary sizes (in percentiles), that is, CDI percentiles decreased by 0.50 when the vowel category variability increased by one standard deviation of the sample mean with the other factors being kept at an average. As a complementary analysis, we provide a correlation matrix and a correlation network plot with all acoustic measures in Appendix 7A and 7B.

Table 5. ***Beta-regression results for vocabulary by parents' input in IDS (n = 21 parent-toddler dyads)***

Parameter	*estimate*	*SE*	z	*p*
Intercept	-0.487	0.225	-2.168	.030*
Pitch_IDS_z	0.266	0.256	1.040	.298
Pitch range_IDS_z	0.051	0.254	0.202	.840
Vowel duration_IDS_z	-0.160	0.274	-0.585	.559
Vowel space_full_IDS_z	0.296	0.264	2.121	.262
Vowel category variability_IDS_z	-0.499	0.254	-1.962	.050*

*$p < .05$

[11] Fitting the pre-registered model resulted in high VIFs for vowel category distinctiveness (VIF = 2.96), vowel space_æ (VIF = 8.34) and vowel space_full (VIF = 9.62). We chose to keep the latter of the vowel space measures, given that this would maximise the information about parents' vowel space.

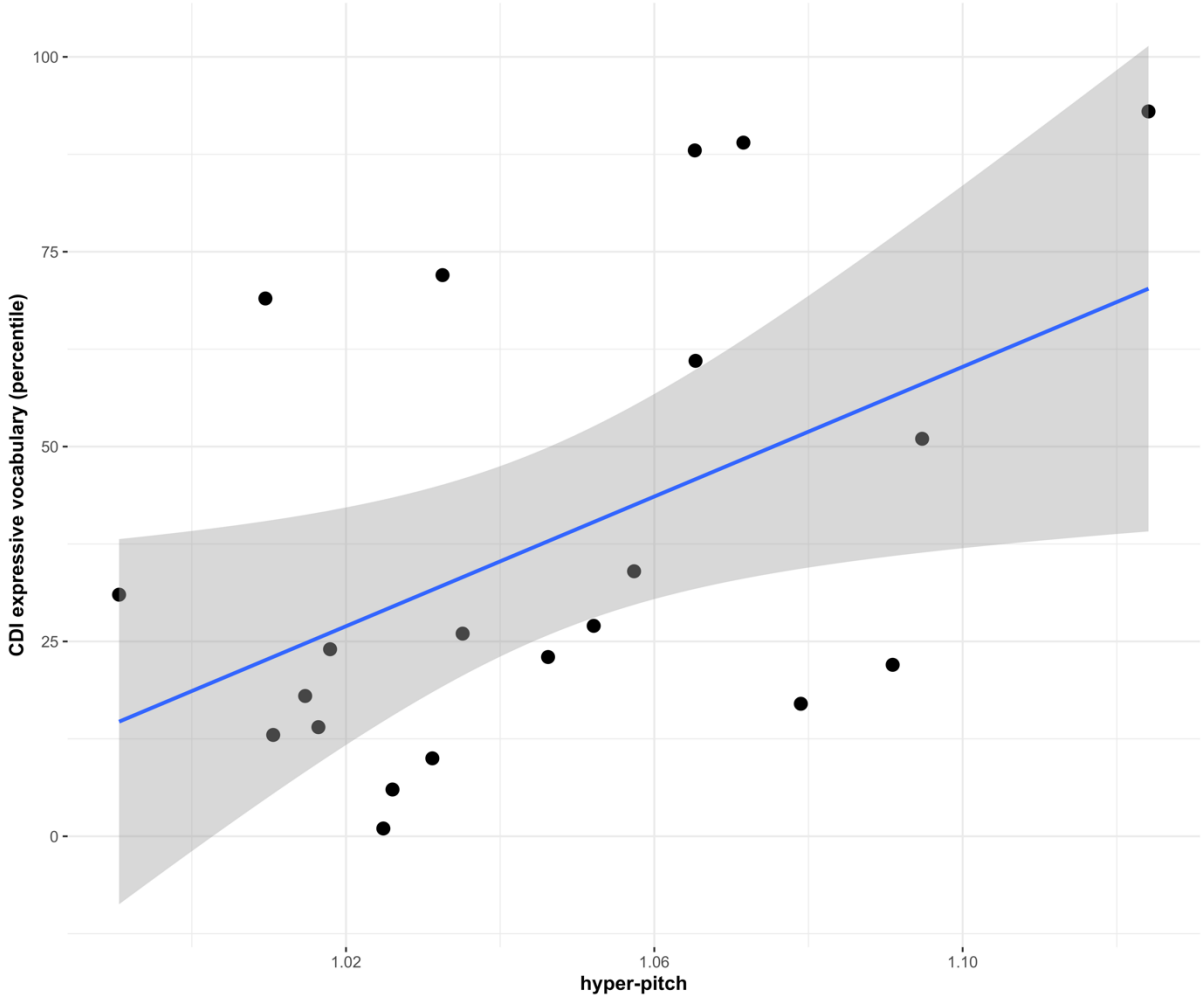


Figure 3. ***Relationship between parents' hyper-pitch and toddlers' vocabulary. Note that the figure visualises the regression line, with the shaded area depicting 95% confidence intervals. Hyper-pitch is the within-parent difference ratio of average pitch, in semitones, in IDS vs ADS.***

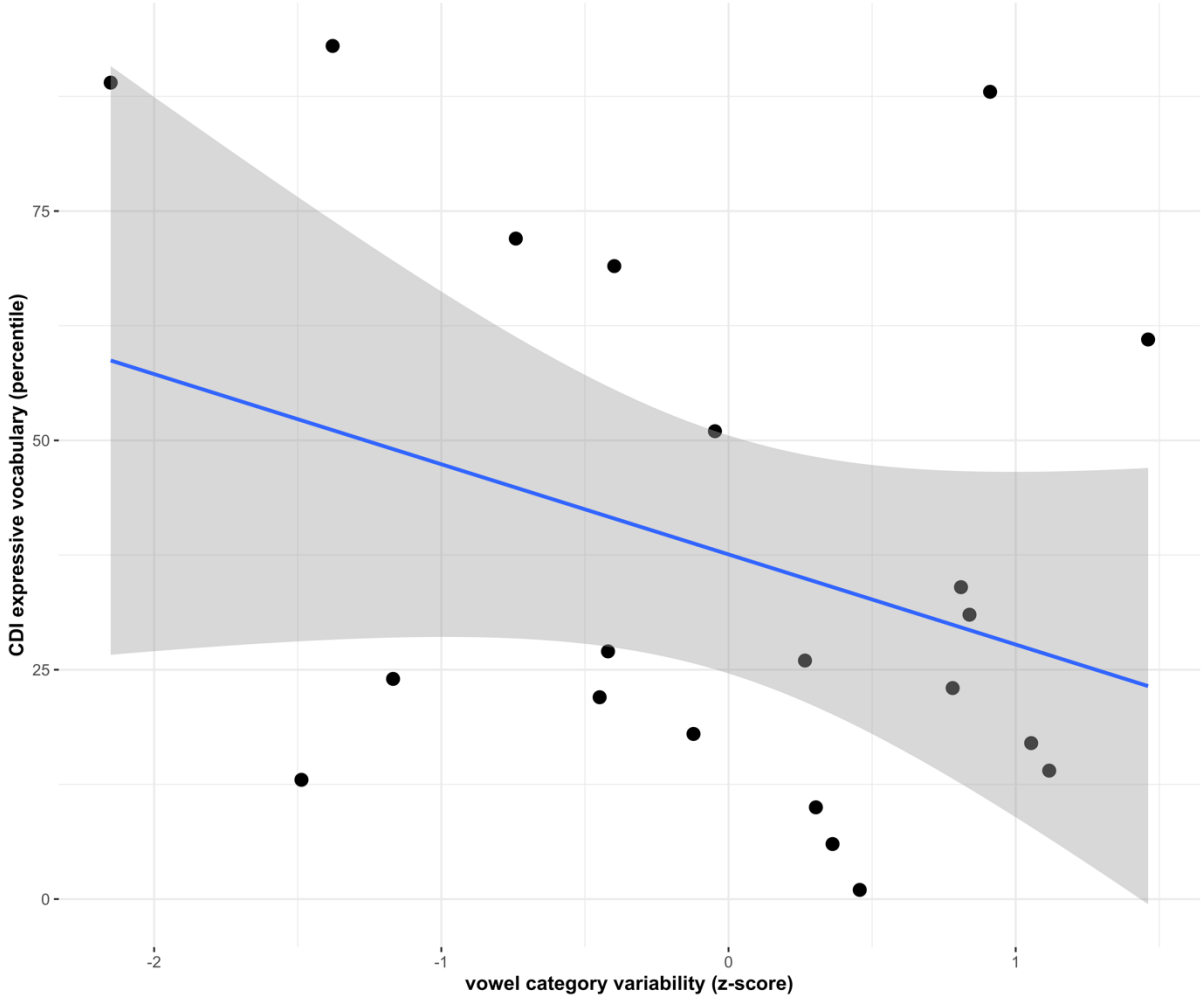


Figure 4. ***Relationship between parents' vowel category variability in IDS and toddlers' vocabulary. Note that the figure visualises the regression line, with the shaded area depicting 95% confidence intervals. The X-axis represents the z-scaled (within mothers and fathers) category variability.***

Discussion

The current study aimed to expand the knowledge about IDS in understudied languages and its potentially facilitating role in early language development. To achieve these aims, we undertook three steps: (1) examined speech of Norwegian parents speaking a Northern Norwegian dialect to their 18-month-old toddlers by measuring traditionally reported and novel acoustic properties of IDS and their differences with respect to ADS; (2) assessed the role of within-parent adaptation between IDS and ADS in predicting toddlers' expressive vocabulary, and, finally, (3) assessed the role of

acoustic properties in IDS itself, in predicting toddlers' expressive vocabulary. Both IDS and ADS were elicited via a storybook reading, to control for within and between-parent differences in linguistic context that can affect speech production.

Overall, the results of the current study, in Norwegian parents to 18-month-old toddlers, supported the first hypothesis on speech 'adaptation' in IDS, as compared to ADS, providing further evidence to the growing body of research indicating that the speech register we use when interacting with young children has unique features, also in a language and a dialect that uses pitch and duration as lexical cues. Parents in our sample had higher mean phrasal pitch, wider phrasal pitch range, and longer vowel durations in IDS over ADS, although the latter was only true for mothers and not fathers. These results are in line with previous studies in other Norwegian dialects (Englund & Behne, 2005, 2006; Kartushina et al., 2021); yet, the gender differences in vowel duration suggest that fathers might be more restrained in IDS than mothers, which goes against the hypotheses that fathers' more energetic interaction style, as compared to mothers, is also manifested in IDS acoustics (Benders et al., 2021). Still, fathers in our study increased their pitch range in IDS, and thus the lack of vowel prolongation could also be related to our limited sample size for fathers, cross-linguistic differences and/or task demands, that is, a storybook reading. Further, parents expanded their vowel space area in IDS more than in ADS, both when examining the corner vowels that are typically reported in the literature (/i:/, /**ɑ**:/, /u:/), the corner vowels particular to the Norwegian language (/i:/, /**æ**:/, /u:/), and the full vowel space covering all border vowels in Norwegian (/ɑ:/, /e:/, /i:/, /u:/, /ʉ:/, /æ:/, /ɔ:/). This result is consistent with the studies in English (Kalashnikova & Burnham, 2018), Russian and Swedish (Kuhl et al., 1997; Marklund & Gustavsson, 2020), Spanish and Basque (Kalashnikova & Carreiras, 2021), as well as Eastern Norwegian (Kartushina et al., 2021), but not Central Norwegian (Englund & Behne, 2006). Apart from differences in the methodologies between the current and Englund and Behne's study, differences in the results on vowel space expansion between these two studies can be attributed either to fine-grained variations within a language (due to dialectal differences), or to differences in children's ages (0–6-month-old infants in Englund and Behne's study). However, vowel categories were more variable in IDS, suggesting that vowel space expansion did not necessarily translate into more intelligible speech. This supports previous work showing more variable underlying vowel categories in speech addressed to infants and toddlers (Cristia & Seidl, 2014; Martin et al., 2015; McMurray et al., 2013; Miyazawa et al., 2017). Furthermore, in Norwegian, such variability has been found in speech to 8–9-month-old infants (Kartushina et al., 2021), and now to 18-month-old toddlers, suggesting no changes in variability with the child's age. The 'sloppiness' of vowel production in IDS could potentially be a side effect of a larger vowel space expansion, or increased pitch variability, that impacts both F1 and F2 (McMurray et al., 2013). Finally, vowel category distinctiveness was comparable across registers, suggesting that although the vowel space was expanded, and the variability of individual vowel categories was increased in parents' IDS, the vowel type

did not appear less identifiable within the participants' vowel clusters across registers. This could be due to parents taking extra care due to the rich vowel inventory of Norwegian, encompassing a total of 19 categories (nine long, nine short, plus schwa). Future work should expand on this result by assessing a bigger range of vowel tokens per participant, and preferably in other languages and dialects that have closer or more distributed mappings of their vowels in F1-F2 space.

With respect to our second hypothesis on the role of differences between IDS and ADS in early language development, our results showed that parents' hyper-pitch predicted toddlers' vocabulary, whereas the other acoustic measures included in our model did not. In other words, those parents who exaggerated their average pitch to a greater degree when reading to their toddlers (as compared to an experimenter), had toddlers with larger vocabulary sizes. Experimental studies have similarly highlighted the role of pitch, in supporting word segmentation in 9-month-old infants (Schreiner & Mani, 2017), and word learning in older toddlers (Graf Estes & Hurley, 2013). Recall, that increase in pitch has been reported as one of the few acoustic features present in the majority of the examined studies, suggesting it to be one of the most salient cues in IDS. In addition, research has shown that infants display larger preference for IDS at older ages (The ManyBabies Consortium, 2020), and this preference is suggested to be driven mainly by pitch increase (Segal & Newman, 2015). Thus, such a preference might engage parents in using higher pitch when interacting with their toddlers, as toddlers might be more responsive in return. As Norwegian uses pitch accent as both a lexically contrastive cue and a cue to mark dialects, parents' pitch increase, as shown in the current study, might also help toddlers incorporate these cues, thus scaffolding the development of their vocabulary.

Finally, with respect to our third hypothesis that addressed the role of direct acoustic infant-directed input in early language development, vowel category variability correlated negatively with toddlers' expressive vocabulary. This result suggests that input containing more precise vowels with little variability within each vowel category may provide scaffolding cues to build a richer vocabulary as reliable vowel productions would facilitate phonological discrimination and establishment of more stable phonological representations (see e.g., Bosch & Ramon-Casas, 2011; Bosch & Ramon-Casas, 2009; Cristia, 2011), facilitating, in turn, the vocabulary acquisition. Although laboratory studies have found facilitatory effects of vowel space expansion on speech processing (Peter et al., 2016; Song et al., 2010), experimental stimuli are *de-facto* less variable, and thus, compact categories might play more important role in 'real life' input, as compared to an experimental setting. Our result is in contrast with that of Hartman and colleagues (2017), who found that vowel space area in IDS, and not vowel variability, predicted vocabulary in similar aged English-learning toddlers. This discrepancy in the results could be due to cross-linguistic differences in vowel realization and variability and/or to differences in the number of analysed vowels; note that Hartman and colleagues examined the three corner vowels only, which

might not have captured parents' full vowel inventory, as attempted in the current study with all Norwegian long vowels.

Crucially, our study demonstrates that the properties of IDS that relate to language outcomes might depend on whether the IDS is operationalised as the acoustic input directed towards the child, or as a within-parent perceptual adaptation when addressing their speech to a child as compared to an adult, respectively. It might be that hyper-pitch as a predictor of vocabulary does not reflect benefits of the acoustic signal *per se*, but rather parents' investment in capturing the attention of their toddlers, and thus such hyper-measures might be better thought of as an index of engagement and parenting style, rather than barely an acoustic booster. Although we did not find any relationship between parents' attitudes towards book reading and the quality of the linguistic input in early childhood and their degree of hyper-pitch, these were exploratory analyses and were not necessarily suited to untangle such a relationship. On the contrary, parents' precision in vowel production when interacting with their children, regardless of the differences with the ADS, correlated with their toddler's vocabulary size. Within this framework it seems more plausible to suggest benefits directly related to the acoustic signal of speech itself. Yet, we note that both of these findings are purely correlational. We need to acknowledge that third variables, such as the time parents spent with the child, or the SES – lacking diversity in our sample – might be mediating these relationships. It has also been suggested that linguistic input has the best function when it is tailored to and matches the linguistic level of the child (Rowe & Snow, 2020). Precisely, recent studies suggest that parents are experts in tuning their speech to their toddlers' needs, both lexically (Leung et al., 2021), but also acoustically (Han et al., 2020, 2021). As such, vocabulary size and parent input might be bi-directional in nature: Toddlers with a richer vocabulary (as opposed to poor) may encourage parents to increase their engagement during storybook reading more (i.e., with hyper-pitch), which, in turn, can lead to clearer (engaging, scaffolding) input to the child.

The current study has several limitations that can be addressed in future research. First, given that we did not target mothers and fathers specifically, but asked the primary caregiver to come to the lab, fathers were underrepresented in our sample, not allowing us to evaluate parent gender differences in IDS more systematically, which have been illustrated elsewhere (Benders et al., 2021). Given that Norway is a highly egalitarian society, where fathers, through the social policy, are promoted as equally important and invested caregivers with the same number of weeks of parental leave as mothers (Brandth & Kvande, 2020), this should be further investigated. Second, we used parent-reported vocabulary as our outcome measure, and although the CDI has shown to be convergent with direct child-based measures of word comprehension (Lo et al., 2021), there is a need to connect properties of IDS with direct language measures in children. Finally, the current study only captured a particular moment in time, and as parents' IDS might change across development (Narayan &

McDermott, 2016), and thus exercise varying influence on language outcomes (McMurray et al., 2013; Rowe & Snow, 2020), longitudinal studies that depict these trajectories would provide stronger evidence of such trajectories over time.

In sum, the current study provides evidence that IDS to 18-month-old Norwegian toddlers follows the same prosodic characteristics as typically reported in the literature for other languages, including increased pitch, pitch range, vowel duration (for mothers), as well as vowel space expansion, although previously reported absent in Norwegian parents to 6-month-olds (Englund, 2018). Yet, additional analyses revealed that parents' vowel categories were more variable in IDS than ADS, in line with previous research, providing evidence that parental vowel categories in IDS are less consistent and more overlapping than in ADS. Furthermore, our study indicates that hyper-pitch as well as low vowel category variability in IDS were positively associated with toddlers' vocabulary. Although the direction and the cause of the effects cannot be asserted with our design, this suggests that parents' increase in pitch when interacting with their child and their consistency in vowel production may facilitate early word learning.

References

Audibert, N., & Falk, S. (2018). Vowel space and f0 characteristics of infant-directed singing and speech. *Proceedings of the International Conference on Speech Prosody*, 153–157. https://doi.org/10.21437/SpeechProsody.2018-31

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development, 36*(4), 847–862. https://doi.org/10.1016/j.infbeh.2013.09.001

Benders, T., StGeorge, J., & Fletcher, R. (2021). Infant-directed speech by Dutch fathers: Increased pitch variability within and across utterances. *Language Learning and Development, 0*(0), 1–34. https://doi.org/10.1080/15475441.2021.1876698

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer [Computer program]*. http://www.praat.org/

Bosch, L., & Ramon-Casas, M. (2009). Phonetic variability in bilinguals' acquisition of native-vowel category contrasts. *The Journal of the Acoustical Society of America, 125*(4), 2770–2770. https://doi.org/10.1121/1.4784720

Bosch, L., & Ramon-Casas, M. (2011). Variability in vowel production by bilingual speakers: Can input properties hinder the early stabilization of contrastive categories? *Journal of Phonetics, 39*(4), 514–526. https://doi.org/10.1016/j.wocn.2011.02.001

Brandth, B., & Kvande, E. (2020). *Designing Parental Leave Policy: The Norway Model and the Changing Face of Fatherhood*. Policy Press.

Broesch, T. L., & Bryant, G. A. (2015). Prosody in infant-directed speech is similar across Western and traditional cultures. *Journal of Cognition and Development, 16*(1), 31–43. https://doi.org/10.1080/15248372.2013.833923

Burnham, E. B., Wieland, E. A., Kondaurova, M. V., McAuley, J. D., Bergeson, T. R., & Dilley, L. C. (2015). Phonetic modification of vowel space in storybook speech to infants up to 2 years of age. *Journal of Speech, Language & Hearing Research, 58*(2), 241–253. https://doi.org/10.1044/2015_JSLHR-S-13-0205

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., Durrant, S., Fennell, C. T., Fiévet, A.-C., Frank, M. C., Gampe, A., Gervain, J., Gonzalez-Gomez, N., Hamlin, J. K., Havron, N., Hernik, M., Kerr, S., Killam, H., Klassen, K., … Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920974622. https://doi.org/10.1177/2515245920974622

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development, 91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*(5), 1584–1595. https://doi.org/10.1111/j.1467-8624.1990.tb02885.x

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software, 34*, 1–24. https://doi.org/10.18637/jss.v034.i02

Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *Journal of the Acoustical Society of America, 129*(5), 3271–3280. https://doi.org/10.1121/1.3562562

Cristia, A. (2013). Input to language: the phonetics and perception of infant-directed speech. *Language and Linguistics Compass, 7*(3), 157–170. https://doi.org/10.1111/lnc3.12015

Cristia, A. (2022). A systematic review suggests marked differences in the prevalence

of infant-directed vocalization across groups of populations. *Developmental Science,* e13265. https://doi.org/10.1111/desc.13265

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language, 41*(4), 913–934. https://doi.org/10.1017/S0305000912000669

Englund, K. T. (2018). Hypoarticulation in infant-directed speech. *Applied Psycholinguistics, 39*(01), 67–87. https://doi.org/10.1017/s0142716417000480

Englund, K. T., & Behne, D. (2005). Infant directed speech in natural interaction—Norwegian vowel quantity and quality. *Journal of Psycholinguistic Research, 34*(3), 259–280.

Englund, K. T., & Behne, D. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development: An International Journal of Research and Practice, 15*(2), 139–160.

Farran, L. K., Lee, C.-C., Yoo, H., & Oller, D. K. (2016). Cross-cultural register differences in infant-directed speech: An initial study. *PloS One, 11*(3), e0151518. https://doi.org/10.1371/journal.pone.0151518

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development, 60*(6), 1497–1510.

Fernald, A., Taeschner, T., Dunn, J., Papoušek, M., Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16,* 477–501.

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., White, L., Goslin, J., & Vihman, M. (2016). British English infants segment words only with exaggerated infant-directed speech stimuli. *Cognition, 148,* 1–9. https://doi.org/10.1016/j.cognition.2015.12.004

Fox, J., & Weisberg, S. (2018). *An R Companion to Applied Regression*. SAGE Publications.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data*. *Journal of Child Language, 44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

García-Sierra, A., Ramírez-Esparza, N., Wig, N., & Robertson, D. (2021). Language learning as a function of infant directed speech (IDS) in Spanish: Testing neural

commitment using the positive-MMR. *Brain and Language, 212,* 104890. https://doi.org/10.1016/j.bandl.2020.104890

Garnier, M., Ménard, L., & Alexandre, B. (2018). Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? *The Journal of the Acoustical Society of America, 144*(2), 1059–1074. https://doi.org/10.1121/1.5051321

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science, 24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy, 18*(5), 797–824. https://doi.org/10.1111/infa.12006

Han, M., de Jong, N. H., & Kager, R. (2020). Pitch properties of infant-directed speech specific to word-learning contexts: A cross-linguistic investigation of Mandarin Chinese and Dutch. *Journal of Child Language, 47*(1), 85–111. https://doi.org/10.1017/S0305000919000813

Han, M., de Jong, N. H., & Kager, R. (2021). Language specificity of infant-directed speech: speaking rate and word position in word-learning contexts. *Language Learning and Development, 17*(3), 221–240. https://doi.org/10.1080/15475441.2020.1855182

Hartman, K. M., Ratner, N. B., & Newman, R. S. (2017). Infant-directed speech (IDS) vowel clarity and child language outcomes. *Journal of Child Language, 44*(5), 1140–1162. https://doi.org/10.1017/S0305000916000520

Hembacher, E., & Frank, M. C. (2020). The early parenting attitudes questionnaire: Measuring intuitive theories of parenting and child development. *Collabra: Psychology, 6*(1). https://doi.org/10.1525/collabra.190

Hirst, D. (2012). *Analyse_tier.praat* [Praat script]. https://www.mail-archive.com/praat-users@yahoogroups.co.uk/msg00061.html

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of Child Language, 45*(5), 1035–1053. https://doi.org/10.1017/S0305000917000629

Kalashnikova, M., Carignan, C., & Burnham, D. (2017). The origins of babytalk: smiling, teaching or social convergence? *Royal Society Open Science, 4*(8), 170306. https://doi.org/10.1098/rsos.170306

Kalashnikova, M., & Carreiras, M. (2021). Input quality and speech perception development in bilingual infants' first year of life. *Child Development, n/a*(n/a). https://doi.org/10.1111/cdev.13686

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology, 5*, 1246. https://doi.org/10.3389/fpsyg.2014.01246

Kartushina, N., Mani, N., Aktan-Erciyes, A., Alaslani, K., Aldrich, N. J., Almohammadi, A., ..., & Mayor, J. (2022). COVID-19 first lockdown as a window into language acquisition: associations between caregiver-child activities and vocabulary gains. *Language Development Research*, *2*(1), 1–36. https://doi.org/10.34842/abym-xv34

Kartushina, N., Robbestad, S., & Mayor, J. (2021, April 7). *The Role of Parental Speech in Infant Language Development: Insights from 8-9-month-old Norwegian Infants*. Society for Research in Child Development 2021 Virtual Biennial Meeting. https://www.srcd.org/event/srcd-2021-biennial-meeting

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition? *First Language*. https://doi.org/10.1177/01427237211066405

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684–686.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.
Leung, A., Tunkel, A., & Yurovsky, D. (2021). Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science*, *32*(7), 975–984. https://doi.org/10.1177/0956797621993104

Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science, 6*(3), F1–F10. https://doi.org/10.1111/1467-7687.00275

Lo, C. H., Rosslund, A., Chai, J. H., Mayor, J., & Kartushina, N. (2021). Tablet assessment of word comprehension reveals coarse word representations in 18–20-month-old toddlers. *Infancy, 26*(4), 596–616. https://doi.org/10.1111/infa.12401

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development, 7*(3), 185–201. https://doi.org/10.1080/15475441.2011.579839

Marklund, E., & Gustavsson, L. (2020). The dynamics of vowel hypo- and hyperarticulation in Swedish infant-directed speech to 12-month-olds. *Frontiers in Communication, 5,* 523768. https://doi.org/10.3389/fcomm.2020.523768

Marklund, E., Marklund, U., & Gustavsson, L. (2021). An association between phonetic complexity of infant vocalizations and parent vowel hyperarticulation. *Frontiers in Psychology, 12*, 2873. https://doi.org/10.3389/fpsyg.2021.693866

Monsen, R. B., & Engebretson, A. M. (1983). The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. *Journal of Speech, Language, and Hearing Research, 26*(1), 89-97. https://doi.org/10.1044/jshr.2601.89

Mæhlum, B., & Røyneland, U. (2012). *Det norske dialektlandskapet*. Cappelen Damm Akademisk.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science, 26*(3), 341–347. https://doi.org/10.1177/0956797614562453

McClay, E. K., Cebioglu, S., Broesch, T., & Yeung, H. H. (2021). Rethinking the phonetics of baby-talk: Differences across Canada and Vanuatu in the articulation of mothers' speech to infants. *Developmental Science, n/a*(n/a), e13180. https://doi.org/10.1111/desc.13180

McCloy, D. (2016). *PhonR: tools for phoneticians and phonologists* (1.0-7) [Computer software]. https://cran.r-project.org/web/packages/phonR/phonR.pdf

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition, 129*(2), 362–378. https://doi.org/10.1016/j.cognition.2013.07.015

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition, 166,* 84–93. https://doi.org/10.1016/j.cognition.2017.05.003

Narayan, C. R., & McDermott, L. C. (2016). Speech rate and pitch characteristics of

infant-directed speech: Longitudinal and cross-linguistic observations. *The Journal of the Acoustical Society of America, 139*(3), 1272–1281. https://doi.org/10.1121/1.4944634

Peter, V., Kalashnikova, M., Santos, A., & Burnham, D. (2016). Mature neural responses to infant-directed speech but not adult-directed speech in pre-verbal infants. *Scientific Reports, 6*(1), 1–14. https://doi.org/10.1038/srep34273

Porritt, L. L., Zinser, M. C., Bachorowski, J.-A., & Kaplan, P. S. (2014). Depression diagnoses and fundamental frequency-based acoustic cues in maternal infant-directed speech. *Language Learning and Development, 10*(1), 51–67. https://doi.org/10.1080/15475441.2013.802962

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. www.R-project.org

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language, 47*(1), 5–21. https://doi.org/10.1017/S0305000919000655

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M.-C., & Cohen, D. (2013). Motherese in interaction: At the cross-road of emotion and cognition? (A systematic review). *PloS One, 8*(10), e78103–e78103. https://doi.org/10.1371/journal.pone.0078103

Schreiner, M. S., & Mani, N. (2017). Listen up! Developmental differences in the impact of IDS on speech segmentation. *Cognition, 160*, 98–102. https://doi.org/10.1016/j.cognition.2016.12.003

Segal, J., & Newman, R. S. (2015). Infant preferences for structural and prosodic properties of infant-directed speech in the second year of life. *Infancy, 20*(3), 339–351. https://doi.org/10.1111/infa.12077

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian communicative development inventories: Reliability, main developmental trends and gender differences. *First Language, 34*(1), 3–23. https://doi.org/10.1177/0142723713510997

Song, J. Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *The Journal of the Acoustical Society of America, 128*(1), 389–400. https://doi.org/10.1121/1.3419786

Steinlen, A., & Bohn, O. (1999). Acoustic studies comparing Danish vowels, British English vowels, and Danish-accented British English vowels. *The Journal of the*

Acoustical Society of America, 105(2), 1097–1097. https://doi.org/10.1121/1.425143

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science, 20*(6), e12456. https://doi.org/10.1111/desc.12456

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Thiessen, E., Hill, E., Saffran, J., & Thiessen, E. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*(1), 53–71.

Wang, Y., Seidl, A., & Cristia, A. (2015). Acoustic-phonetic differences between infant- and adult-directed speech: The role of stress and utterance position. *Journal of Child Language, 42*(4), 821–842. https://doi.org/10.1017/S0305000914000439
Wichmann, S. (2019). How to distinguish languages and dialects. *Computational Linguistics, 45*(4), 823–831. https://doi.org/10.1162/coli_a_00366

Weirich, M., & Simpson, A. (2019). Effects of gender, parental role, and time on infant- and adult-directed read and spontaneous speech. *Journal of Speech, Language, and Hearing Research, 62*(11), 4001–4014. https://doi.org/10.1044/2019_JSLHR-S-19-0047

Xu Rattanasone, N., Burnham, D., & Reilly, R. G. (2013). Tone and vowel enhancement in Cantonese infant-directed speech at 3, 6, 9, and 12 months of age. *Journal of Phonetics, 41*(5), 332–343. https://doi.org/10.1016/j.wocn.2013.06.001

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Data, Code and Materials Availability Statement

Data, code and materials are available at the Open Science Framework project's page, at the following permanent link: https://osf.io/7st6w/.

Ethics Statement

The current study was conducted according to the guidelines laid down in the Declaration of Helsinki, with written informed consent obtained from a parent or a guardian for a child before any assessment or data collection. The study has been approved

by the Norwegian Centre for Research Data (NSD, ref. 56312), and the local ethical committee at the Department of Psychology, University of Oslo.

Authorship and Contributorship Statement

AR, JM, GÓ and NK conceptualised the study. JM, GÓ and NK supervised the study. AR analysed the data and wrote the first draft of the manuscript. AR, JM, GÓ and NK interpreted the results and revised the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We thank Mathea Sandholm, Hanne Akseth Ulriksen, Madelene Halvari Niska, and Annie-Justicia Karlsson for their help with data collection, and Roger Mundry for help with the vowel category distinctiveness measure. We would like to thank reviewers for their insightful comments and suggestions. AR and NK were supported by the Research Council of Norway through project number 301625, and its Centres of Excellence funding scheme [project number 223265].

Appendices

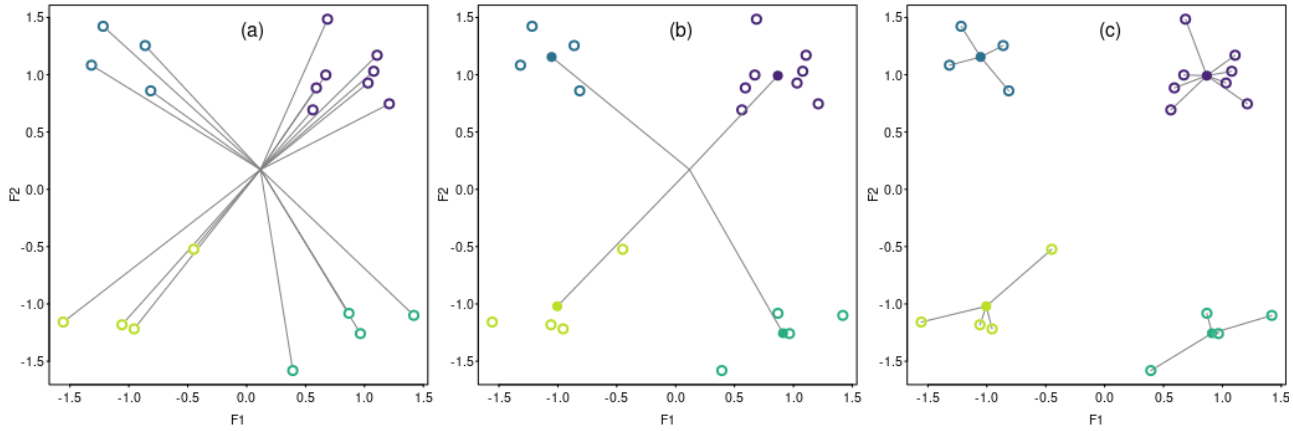
Appendix 1. ***Overview of words in the storybook for each target vowel (in bold)***

/i:/	/y:/	/e:/	/ø:/	/æ:/	/ʉ:/	/u:/	/ɔ:/	/ɑ:/
b**i**l (car)	l**y**s (light)	s**e** (look)	br**ø**d (bread)	d**er** (there)	l**u**e (hat)	b**o**k (book)	s**o**ve (sleep)	ban**a**n (banana)
gr**i**s (pig)	fl**y** (airplane)	skj**e** (spoon)	sn**ø** (snow)	h**er** (here)	p**u**te (pillow)	sk**o** (shoe)	t**o**g (train)	b**a**de (bath)
sp**i**se (eat)	d**y**ne (duvet)	m**e**r (more)	d**ø**r (door)	v**æ**re (be)	k**u** (cow)	f**o**t (foot)	h**å**r (hair)	k**a**ke (cake)
sk**i**ve (slice)	d**y**r (animal)	n**e**se (nose)	bj**ø**rn (bear)	b**æ**re (carry)	m**u**s (mouse)	s**o**l (sun)	m**å**ne (moon)	m**a**ge (belly)
v**i** (we)	n**y** (new)	l**e**se (read)	l**ø**pe (run)	skj**æ**re (cut)	f**u**gl (bird)	hall**o** (hello)	g**å** (go)	br**a** (good)

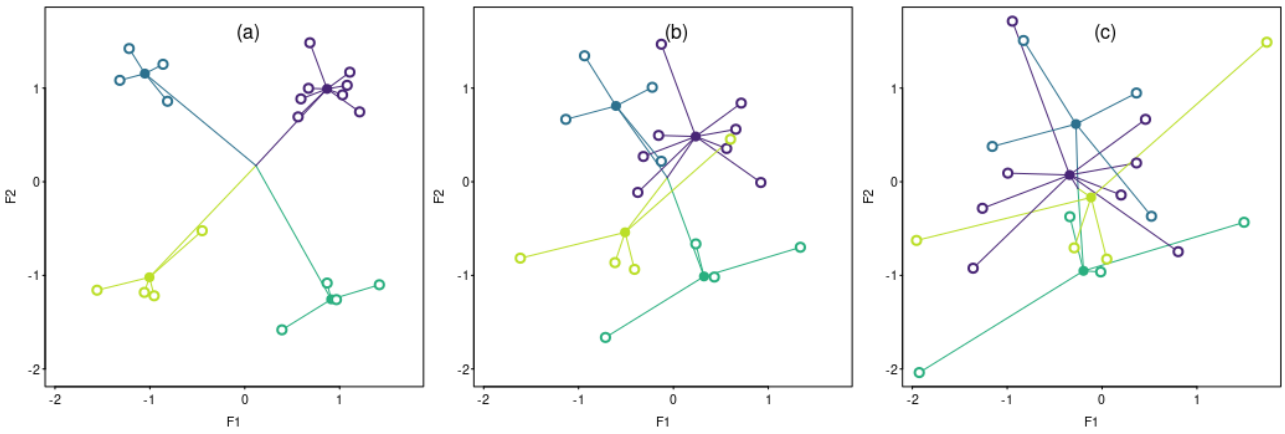
Appendix 2. *Comparison of two 'born-early' dyads to the rest of the sample*

Variable	Not 'born early' (n = 19)	'Born early' (n = 2)	Full sample (n = 21)
Cat_dist.mean.IDS			
Mean (SD)	0.885 (0.0500)	0.903 (0.0203)	0.887 (0.0480)
Median [Min, Max]	0.900 [0.745, 0.952]	0.903 [0.889, 0.918]	0.900 [0.745, 0.952]
Cat_var.mean.IDS			
Mean (SD)	317000 (132000)	249000 (78200)	311000 (128000)
Median [Min, Max]	279000 [109000, 551000]	249000 [194000, 304000]	279000 [109000, 551000]
Vowel_space_full_IDS			
Mean (SD)	433 (106)	521 (193)	441 (113)
Median [Min, Max]	428 [223, 615]	521 [385, 658]	428 [223, 658]
Vowel_space_æ_IDS			
Mean (SD)	433 (106)	521 (193)	441 (113)
Median [Min, Max]	428 [223, 615]	521 [385, 658]	428 [223, 658]
Duration.mean.IDS			
Mean (SD)	132 (21.8)	150 (23.7)	134 (22.0)
Median [Min, Max]	128 [96.0, 187]	150 [133, 167]	128 [96.0, 187]
Pitch_range.mean.IDS			
Mean (SD)	14.9 (3.41)	12.4 (4.55)	14.6 (3.47)
Median [Min, Max]	14.7 [6.41, 22.2]	12.4 [9.21, 15.7]	14.7 [6.41, 22.2]
Pitch.mean.IDS			
Mean (SD)	53.8 (6.38)	54.7 (2.10)	53.9 (6.08)
Median [Min, Max]	56.5 [40.7, 61.4]	54.7 [53.2, 56.2]	56.2 [40.7, 61.4]
Cat_dist_effort			
Mean (SD)	1.00 (0.0569)	1.00 (0.0254)	1.00 (0.0543)
Median [Min, Max]	1.01 [0.859, 1.10]	1.00 [0.985, 1.02]	1.01 [0.859, 1.10]
Cat_var_effort			
Mean (SD)	1.20 (0.373)	1.28 (0.00276)	1.21 (0.355)
Median [Min, Max]	1.10 [0.641, 2.08]	1.28 [1.28, 1.28]	1.15 [0.641, 2.08]
Vowel_space_full_effort			
Mean (SD)	1.15 (0.269)	1.53 (0.760)	1.19 (0.327)
Median [Min, Max]	1.13 [0.667, 1.90]	1.53 [0.992, 2.07]	1.13 [0.667, 2.07]
Vowel_space_æ_effort			
Mean (SD)	1.15 (0.268)	1.23 (0.146)	1.15 (0.257)
Median [Min, Max]	1.07 [0.758, 1.73]	1.23 [1.13, 1.34]	1.09 [0.758, 1.73]
Duration_effort			
Mean (SD)	1.18 (0.162)	1.07 (0.195)	1.17 (0.163)
Median [Min, Max]	1.15 [0.945, 1.65]	1.07 [0.935, 1.21]	1.15 [0.935, 1.65]
Pitch_range_effort			
Mean (SD)	1.12 (0.239)	1.16 (0.315)	1.13 (0.238)
Median [Min, Max]	1.17 [0.655, 1.47]	1.16 [0.935, 1.38]	1.17 [0.655, 1.47]
Pitch_effort			
Mean (SD)	1.05 (0.0341)	1.02 (0.00229)	1.05 (0.0338)
Median [Min, Max]	1.05 [0.991, 1.12]	1.02 [1.01, 1.02]	1.04 [0.991, 1.12]
CDI_prod_percentile			
Mean (SD)	39.3 (30.3)	21.0 (4.24)	37.6 (29.3)
Median [Min, Max]	27.0 [1.00, 93.0]	21.0 [18.0, 24.0]	26.0 [1.00, 93.0]

Appendix 3A. ***Illustration of the method to determine cluster distinctiveness. The total sum of squares (SStot) is the sum of the squared distances of the individual vocalizations from the overall centroid (a). The between cluster sum of squares (SSbetween) is sum of the squared distances of the per-vowel centroids times the number vocalizations per vowel (b). The within cluster sum of squares is the sum of the squared distances of the individual vocalizations from the respective vowel's centroid (c). Each vowel is depicted by a specific color, individual vocalizations by open dots, and vowel centroids by filled dots. Note the SStot = SSbetween + SSwithin. In case of the example SStot = 43.544, SSbetween = 40.609, SSwithin = 2.936, and cluster distinctiveness = 0.933.***



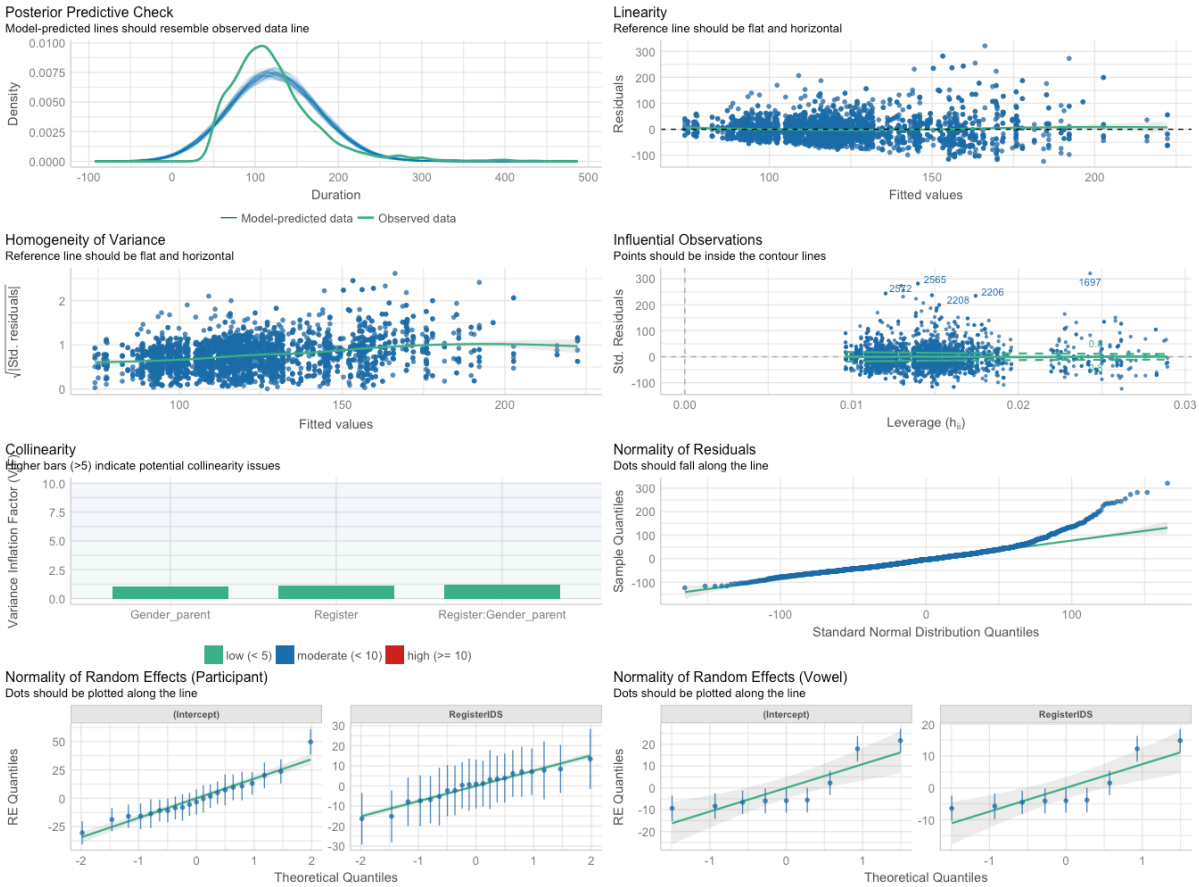
Appendix 3B. ***Illustration of various values of cluster distinctiveness. Each vowel is depicted by a different color, open dots show the individual utterances, and filled dots the clusters' centroids. The total variance explained by vowel type ('cluster distinctiveness') is 0.93 in (a), 0.53 in (b), and 0.14 in (c).***



Appendix 4. ***sessionInfo() output providing R libraries and their versions that were used in the analyses***

Package	Loaded version	Date
betareg	3.1-4	2021-02-09
car	3.0-12	2021-11-06
carData	3.0-4	2020-05-22
doBy	4.6.11	2021-07-13
dplyr	1.0.7	2021-06-18
effsize	0.8.1	2020-10-05
emmeans	1.7.1-1	2021-11-29
emuR	2.3.0	2021-06-11
factoextra	1.0.7	2020-04-01
forcats	0.5.1	2021-01-27
ggplot2	3.3.5	2021-06-25
ggpubr	0.4.0	2020-06-27
ggstatsplot	0.9.0	2021-10-19
knitr	1.36	2021-09-29
lme4	1.1-27.1	2021-06-22
lmerTest	3.1-3	2020-10-23
lsmeans	2.30-0	2018-11-02
Matrix	1.3-4	2021-06-01
patchwork	1.1.1	2020-12-17
performance	0.8.0	2021-10-01
phonR	1.0-7	2016-08-25
purrr	0.3.4	2020-04-17
qqplotr	0.0.5	2021-04-23
rcompanion	2.4.6	2021-11-21
readr	2.1.1	2021-11-30
readxl	1.3.1	2019-03-13
shinyBS	0.61	2015-03-31
soundgen	2.5.0	2021-11-21
stringr	1.4.0	2019-02-10
table1	1.4.2	2021-06-06
tibble	3.1.6	2021-11-07
tidyr	1.1.4	2021-09-27
tidyverse	1.3.1	2021-04-15
viridis	0.6.2	2021-10-13
viridisLite	0.4.0	2021-04-13
vowels	1.2-2	2018-03-05

Appendix 5A. *Model diagnostics of vowel duration pre log-transformation*

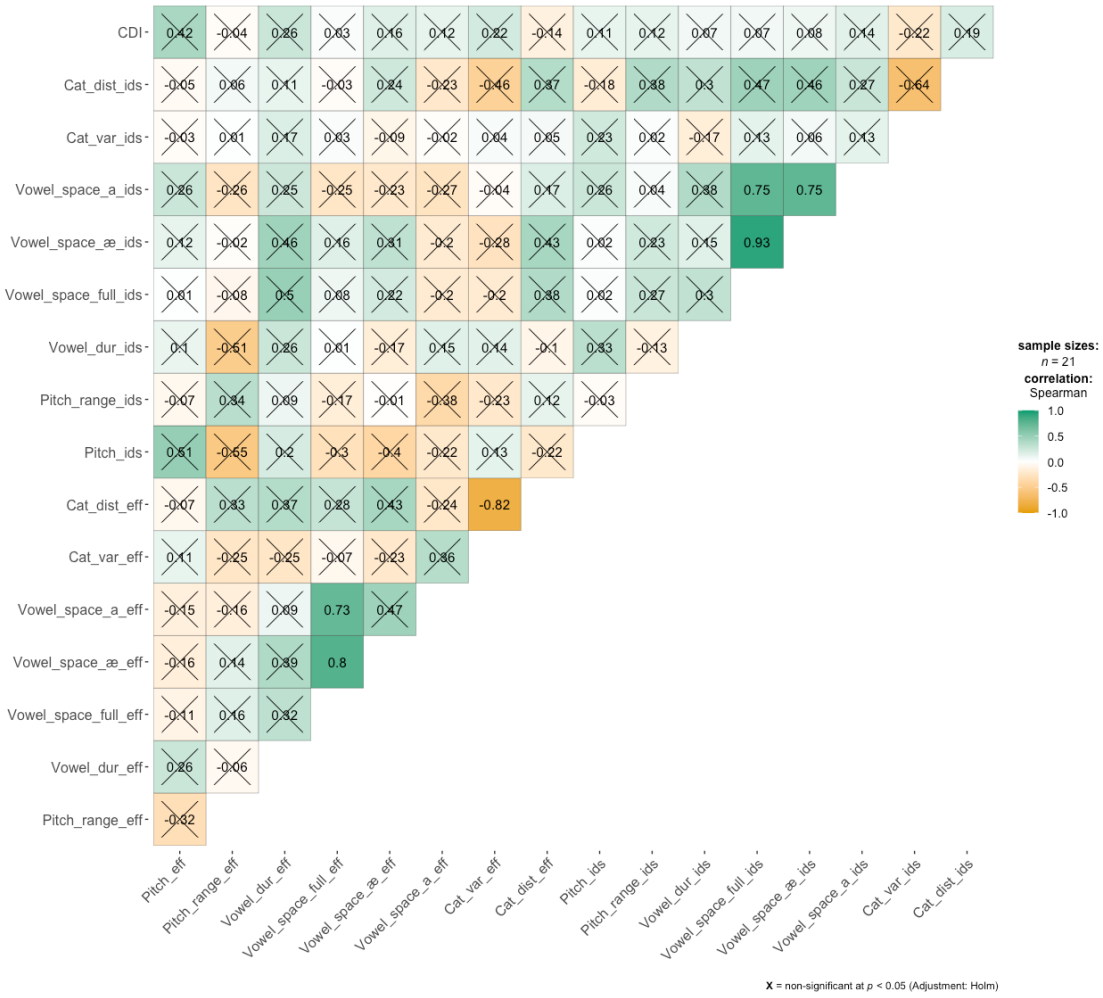


Appendix 5B. ***Model diagnostics of vowel duration post log-transformation***

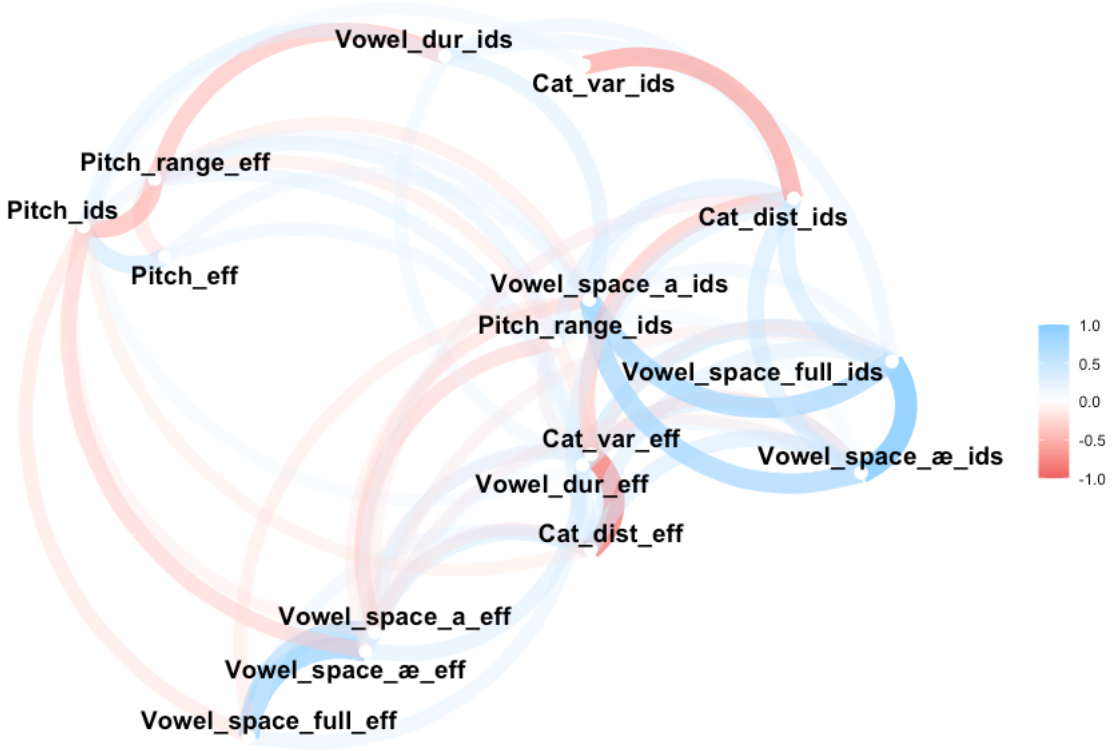
Appendix 6A. *Model diagnostics of vowel category variability pre log-transformation*

Appendix 6B. ***Model diagnostics of vowel category variability post log-transformation***

Appendix 7A. *Spearman correlation matrix of difference (_eff) and IDS-input (_ids) acoustic measures*



Appendix 7B. ***Spearman correlation network of difference (_eff) and IDS-input (_ids) acoustic measures*****. Note that only correlations stronger than +/- .20 are displayed**



License

Large-scale study of speech acts' development in early childhood

Mitja Nikolaus
Aix Marseille Univ, Université de Toulon, CNRS, LIS, LPL, Marseille, France

Eliot Maes
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Jeremy Auguste
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Laurent Prévot
Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France

Abdellah Fourtassi
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Abstract: Studies of children's language use in the wild (e.g., in the context of child-caregiver social interaction) have been slowed by the time- and resource- consuming task of hand annotating utterances for communicative intents/speech acts. Existing studies have typically focused on investigating rather small samples of children, raising the question of how their findings generalize both to larger and more representative populations and to a richer set of interaction contexts. Here we propose a simple automatic model for speech act labeling in early childhood based on the INCA-A coding scheme (Ninio et al., 1994). After validating the model against ground truth labels, we automatically annotated the entire English-language data from the CHILDES corpus. The major theoretical result was that earlier findings generalize quite well at a large scale. Further, we introduced two complementary measures for the age of acquisition of speech acts which allows us to rank different speech acts according to their order of emergence in production and comprehension. Our model is shared with the community so that researchers can use it with their data to investigate various question related to language use both in typical and atypical populations of children.

Keywords: language acquisition; conversation; language use; speech acts

Corresponding author: Mitja Nikolaus, Aix-Marseille University, LIS, 163 Avenue de Luminy, 13288, Marseille, France. Email: mitja.nikolaus@univ-amu.fr.

ORCID ID: https://orcid.org/0000-0001-5609-6628

Citation: Nikolaus, M., Maes, E., Auguste, J., Prevot, L., & Fourtassi, A. (2022). Large-scale study of speech acts' development in early childhood. *Language Development Research, 2*(1), 268–305. https://doi.org/10.34842/2022.0532

Introduction

Research on language learning has largely focused on investigating how children acquire language form (e.g., phonology, lexicon, and syntax) and content (e.g., word and sentence meanings). Yet, an important aspect of language learning, which has received less attention, is the mastery of how to use language adequately in natural social interactions (Bloom & Lahey, 1978). This mastery involves, in particular, using linguistic utterances to encoding and decode communicative intents (Grice, 1975) or speech acts that characterize the illocutionary force of an utterance (e.g question, assertion, and request) (Searle, 1976). Children's learning of speech acts is crucial for their ability to engage in coherent conversations. For example, it is important to recognize that an utterance is a "question" requiring an "answer", or that it is a "request" requiring "acceptance" or "refusal", instead.

Several taxonomies have been proposed that purport to capture children's emergent repertoire of speech act categories in the context of early child-caregiver social interactions (for reviews, see Cameron-Faulkner, 2014; Casillas & Hilbrink, 2020), the most comprehensive to date is the Inventory of Communicative Acts and its abridged version INCA-A (Ninio et al., 1994).

Snow et al. (1996) used INCA-A to study the emergence of speech act major classes in a longitudinal corpus of children aged 14 to 32 months old.[1] They documented several important findings that not only informed our understanding of language use development, but also shed light on how children's emerging linguistic skills interface with the development of their social-cognitive competences. By analyzing the development of the number of distinct speech acts as well as the distribution of speech acts used by children, they showed that when children utter their first words, they already express a range of simple communicative intents such as requests and questions. The repertoire of speech acts was observed in this study to increase rapidly within the first years of life, in tandem with development in social-cognitive and linguistic skills: Children become able to express more sophisticated speech acts such as "promise", "prohibit", and "persuade". Using the same coding scheme, Rollins (1999, 2017) has shown that investigating speech act development can also help us study atypical cognitive development such as autism.

While this previous effort has been influential in the study of language use development, it has relied on hand annotation to code the data, which has limited the researchers' ability to explore how their findings generalize to larger population of children and across different interactive contexts. In fact, INCA-A is a rather complex scheme with a

[1] While the terms "speech act" and "communicative intent" have sometimes been used by different researchers to mean slightly different things or to refer to different taxonomies, here — and for simplicity — we use them interchangeably to refer to the categories of communicative intents at the utterance level, as defined in the INCA-A coding scheme.

large number of categories (e.g., 67 different types of illocutionary acts) and its hand-annotation — including the effort of train annotators — is prohibitively expensive to deploy at a large scale.

Current study

The current study aims at addressing this gap using recent advances in automatic speech act labeling. Using Snow et al.'s child-caregiver corpus and its INCA-A annotation, we tested various models on their ability to map utterances to corresponding speech acts and we selected the one that provided the best performance on a testing set made of unseen utterances from the same corpus.

Using this model, we examined how previous findings in speech act development generalized at scale. To this end we proceeded in two steps: First, we validated the chosen model by testing its ability to replicate key findings from Snow et al. (1996). More specifically, we reproduce developmental patterns regarding the number of distinct speech acts as well as the distribution of speech acts used by children from 14 to 32 months of age. Second, and after successful validation, we used the model to automatically label the entire North American English-language section of CHILDES (MacWhinney, 2017) and compared the results of this large-scale analysis to the original findings.

Additionally, we proposed methods for quantifying the age of acquisition of a speech act both in terms of production and comprehension. These measures have allowed us to rank different speech acts according to their order of emergence. We first examined this order of emergence with data in Snow et al. (1996), and second, thanks to our automatic labelling tool, we tested how this developmental trajectory generalized across all English language corpora in CHILDES.

The paper is organized as follows. First, we introduce the dataset and provide an overview of models for automatic annotation of speech acts that we evaluated in our study. Further, we define the measures for speech act emergence in production and comprehension. In the results sections we compare the performance of the selected models and present replications the findings of Snow et al. (1996) using automatically generated labels. Additionally, the results contain predicted ages of acquisition for each speech act using both manually-annotated and automatically-annotated data. Finally, we discuss the results in the context of language development in general and point out limitations of the current approach which offer possibilities for future research.

Datasets and Methods

Datasets

New England Corpus. For model training and validation, we use ground-truth labels from the dataset collected by Snow et al. (1996) which is the largest child-caregiver interaction dataset annotated for speech acts. This dataset was collected for a longitudinal study of 52 children aged 14, 20 and 32 months old. Child-caregiver dyads were invited for three sessions that consisted of semi-structured free play. All conversations were recorded, transcribed, and annotated with INCA-A coding scheme. There were 55,941 labelled utterances in total.

English-Language CHILDES. In order to test how findings from Snow et al. (1996) generalize to a larger dataset of children and across different contexts, we use the entire North American English-language subset of CHILDES made of children in the same age range (i.e., between 14 and 32 month old), resulting in 2078 different transcripts totaling 354 children.[2]

INCA-A Coding Scheme

INCA-A is the most comprehensive coding scheme to date that was designed to capture children's emerging speech acts it the context of spontaneous social interaction with a caregiver (Ninio et al., 1994). The coding scheme has two coding tiers: 1) the interchange level that annotates the topic of the conversation (e.g., "discussing a recent event"), and may span multiple utterances, and 2) the illocutionary force level (e.g., "Ask a yes/no question") which is determined at the utterance level. Here, we focus on the illocutionary force. INCA-A has 67 different speech act types, which are grouped into several high-level categories such as directives, declarations, commitments, markings, statements, questions, evaluations, and other vocalizations.[3]

Automatic Classification of Speech Acts

Speech act classification (also referred to as dialogue act tagging in the field of Natural Language Processing) describes the task of annotating utterances in dialogue with their respective speech act category. Given a transcript of a conversation and a speech act coding scheme, each utterance in the transcript is assigned one of the speech acts in the coding scheme (Stolcke et al., 2000).

Early work used Hidden Markov Models to map utterances to speech acts using a set of lexical, collocational, and prosodic cues (Stolcke et al., 2000). Subsequent work has used

[2]For fair comparison, we excluded very short transcripts where the number of children's utterances was less than the minimum number of children's utterances in transcripts of the New England corpus at the same age.

[3]Refer to the appendix for the full list of speech acts.

Recurrent Neural Networks (RNNs) such as Long short-term memory networks (LSTMs) for encoding transcribed utterances in order to leverage the sequential structure of the data (Khanpour et al., 2016). More recent approaches combine hierarchical deep neural network encoders with Conditional Random Field (CRF) decoders (Kumar et al., 2018). While the encoder is aware of relationships between the different utterances of a transcript and thus models dependencies in the *feature space*, the CRF can model transition probabilities in the *label space*. In this way, it can for example learn common adjacency pairs (Schegloff & Sacks, 1973) in conversation, e.g. that questions are usually followed by answers.

Following this brief review, we considered and compared the following models.

Baselines

As this work is the first to propose automatic speech act annotation using the INCA-A coding scheme on child-caregiver conversations, we run several baselines in order to obtain reference performances on this specific task.

Majority Classifier. As a first simple baseline, we consider the majority classifier, which always predicts the most frequent speech act.

Random Forests. We use the reference implementation of a random forests algorithm from scikit-learn (Pedregosa et al., 2011). As features, we provide the model with the speaker (caregiver or child), bag-of-words, part-of-speech tags (that are present in the corpus[4]), and the number of words in the utterance.

Support Vector Machine. Using the same features as for the random forests model, we train and evaluate a linear support vector machine from scikit-learn.

Conditional Random Field

Next, we consider a CRF as annotation model. We hypothesized this model would outperform the baselines thanks to its ability to track transition probabilities in the label space. We use *pycrfsuite*[5] (Okazaki, 2007) to implement the CRF. We extend the set of features used by the baseline models and add bigrams and repetitions (words that are repeated from the previous utterance, as well as the number of repeated words normalized by the utterances length) to provide the model with some context of the previous utterances.[6] The model uses the whole conversation in a transcript to find the most probable sequences of labels using the Viterbi algorithm.

[4]The POS tags in CHILDES were automatically generated using the Morphological Analysis algorithm (MOR; MacWhinney, 2000) which yields a high accuracy rate on CHILDES adult data (above 99%).

[5]https://github.com/scrapinghub/python-crfsuite

[6]In preliminary experiments we tested adding all the exact words of previous utterances as features to the model but observed, if anything, a small degradation in performance.

Hierarchical LSTM + CRF

We further consider a model that is inspired by state-of-the-art speech act annotation models in other domains. More specifically, we implement a hierarchical LSTM encoder combined with a CRF decoder similar to the implementation of Kumar et al. (2018). The encoder processes the utterances within a transcript on two levels. We add a special token representing the speaker identity to the beginning of each utterance. Afterwards, for each utterance, one-hot encodings of the words are passed through word embeddings, and are then encoded using the word-level LSTM. The last hidden representation of this LSTM forms the latent utterance representation, which is then passed into the utterance-level LSTM. This higher-level LSTM processes the utterances sequentially and generates conversation-context-aware representations. The output of each timestep of the utterances LSTM is then passed as features to a CRF, which predicts the corresponding speech act. The model has access to contextualized utterance representations as well as the history of speech acts for the classification task. A high-level overview of the architecture of this model can be found in the appendix (Figure 9).

BERT

Given recent developments in NLP regarding the success of pre-trained contextualized embeddings (Devlin et al., 2018), we additionally test the performance of a model where utterances are encoded using BERT. The success of these models relies on self-attention mechanisms that allow the model to create contextualized representations with long-range dependencies as well as setups in which the encoder is pre-trained on large-scale data before being fine-tuned on the actual task. Here we replace the word-level LSTM of the Hierarchical LSTM + CRF model with a pre-trained publicly available implementation of DistilBERT (Wolf et al., 2020). The weights of BERT are fine-tuned on the task. Details on the hyperparameters of the neural network models can be found in the Appendix.

Measures of Speech Act Emergence

Here we introduce measures of speech acts' age of emergence, both at the level of children's production and comprehension.

Production

By analogy to work in word learning (Braginsky et al., 2016; Goodman et al., 2008), we define the age of acquisition of a speech act in production as the month by which at least 50% of the observed children produce it.[7] More precisely, for each speech act s,

[7] In line with Snow et al. (1996), we consider that a child acquired a speech act if it is produced at least twice at a certain age.

we proceed as follows:

1. For each age in the dataset (i.e., 14, 20 and 32 months), calculate the proportion of children who are producing S at least twice.

2. Perform a logistic regression over these proportions.

3. Measure the age of first production as the age where the logistic regression curve surpasses the value 0.5.

Comprehension

Studying speech act emergence only from a production point of view may underestimate children's pragmatic competence. Thus, we additionally introduce a measure for children's comprehension, which we define as the ability of children to respond to a target speech act in a contingent fashion (e.g., responding to a "yes/no question" with "yes" or "no"). More precisely, for each speech act S, we proceed as follows:

1. Find all utterances produced by the caregivers labelled as S.

2. Find all cases where these utterances are followed by an utterance of the child.

3. For each occurring follow-up utterance, annotate whether its speech act is contingent as a response to S.[8] We manually annotated the contingency of all combinations of speech act categories that appear in the data. Using this annotation, we could label each child utterance that follows a caregiver utterance as either possibly contingent or non-contingent based on the corresponding speech act category. The contingency annotation can be found in the GitHub repository: https://github.com/mitjanikolaus/childes-speech-acts.

4. For each age (14, 20 and 32 months), calculate the proportion of contingent follow-up utterances.

5. Perform a logistic regression over the proportion.[9]

6. Measure the age of comprehension as the age where the logistic regression curve surpasses the value 0.5.

[8] Annotating contingency was done using a binary scale, indicating whether the speech act was *possibly* contingent (1) or clearly non contingent (0). A speech act was considered contingent (1) if it can form a coherent response with respect to the previous speech act, and non contingent (0) otherwise.

[9] We only regard data points where the proportion was calculated over at least 2 examples, i.e. where there were at least two utterances with follow-ups.

Table 1: Accuracy for all models.

Model	**Accuracy**
Majority Classifier	13.44% ($\pm 2.81\%$)
Random Forests	62.81% ($\pm 6.29\%$)
Support Vector Machine	62.42% ($\pm 6.97\%$)
Conditional Random Field	**72.33%** ($\pm 4.23\%$)
Hierarchical LSTM + CRF	69.77% ($\pm 3.70\%$)
+ BERT	68.50% ($\pm 4.29\%$)
Inter-Annotator Agreement	81% to 89%

Results and Analyses

First, we compare performance across all models presented above on the New England corpus. Second, we choose the best performing model and test the extent to which its predicted labels replicate major findings obtained using gold labels from Snow et al. (1996). Finally, we use the model to automatically label the North American section from CHILDES and explore how original findings from Snow et al. (1996) on the emergence of speech acts generalize to this larger dataset.

Comparing Models of Speech Act Labeling

We evaluate our models on the speech act annotations of utterances in the New England corpus (Snow et al., 1996). We employ 5-fold cross validation so that we evaluate (and later utilize in all analyses) only the predicted labels on the parts of the corpus that were not seen by the model in the training phase. To this end, and to obtain labels for the whole New England corpus, we train models on 5 different training sets, always holding out 20% of the data. Then we use each of the trained models to label their respective test sets which together form a set of predicted speech act labels for the whole New England corpus.

We report the mean and standard deviation (based on the five cross-validation runs) of each model's accuracy in Table 1. The majority classifier had a high score given the relatively large label space. This could be explained by the fact the label distribution is heavily skewed (Figure 1). A small set of speech acts are used very frequently while several others are rarely used. As for other baseline models, i.e., random forests and support vector machine, the scores are relatively high despite the fact that they do not have access to the conversation history or dependencies in the label space. Our more sophisticated models (Hierarchical LSTM with and without BERT) did not improve performance much, which could be explained by the lack of large-scale training data. Further, in the case of the BERT-based model, we hypothesize that we do not see any

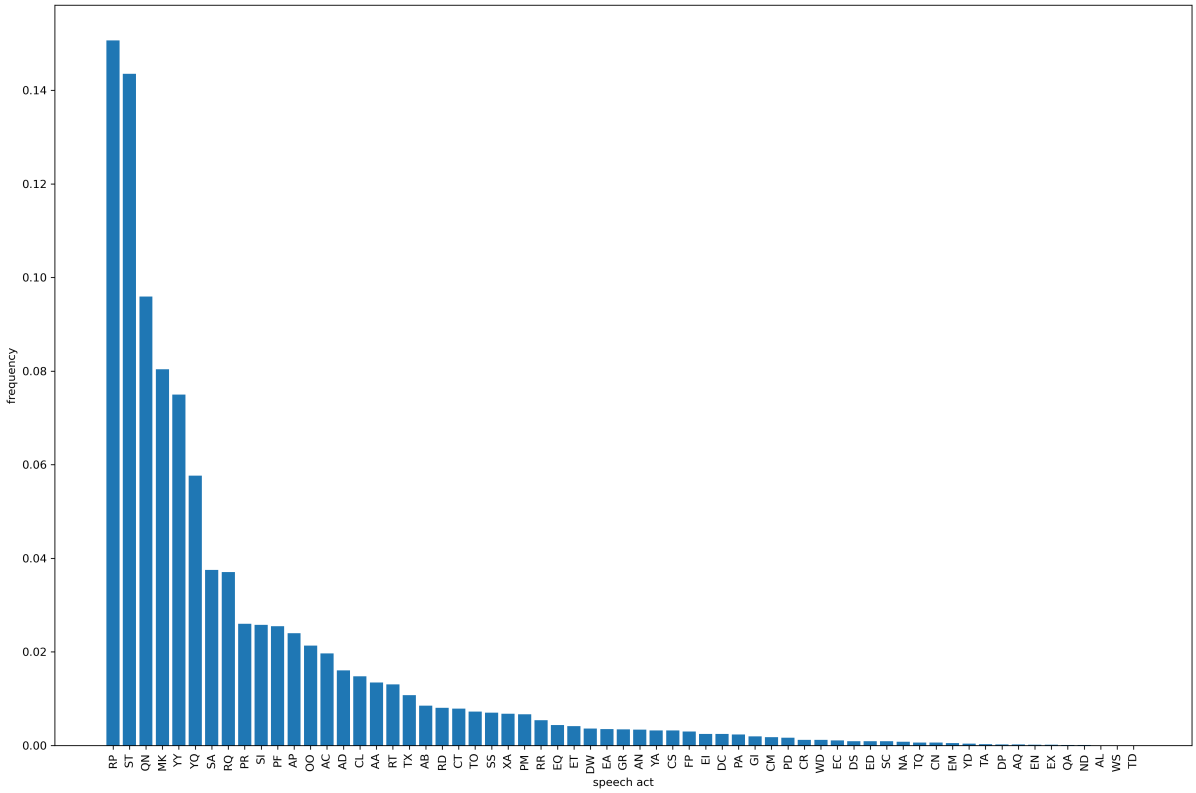


Figure 1. ***Distribution of frequencies of all speech acts in the New England corpus. Labels from the INCA-A tagset are listed in the Appendix.***

performance gains because this model is pre-trained on large text corpora (based on e.g. Wikipedia) that do not have much in common with the dynamics of child-caregiver conversations.

Finally, we find that the CRF model shows the highest accuracy scores, outperforming the baselines as well as the more complex neural network models. Its large performance gains over the baseline are most likely explained by its ability to track transition probabilities in the label space. This property is crucial for the task of speech act annotation; given a speech act sequence, certain speech acts are very likely to follow and others are not. The CRF is the best-performing model, and thus, it is the one we for the rest of analyses in the paper.

Amount of Training Data

We further investigate the effects of the amount of training data on the performance of the CRF model. Figure 2 presents the test accuracy as a function of training set size for this model. The performance indicated in Table 1 was obtained when the model was

trained on 80% of the dataset (around 44,000 utterances). However, from the learning curve in Figure 2 we can see that the model actually achieves decent scores (around 65% accuracy) when trained on only 5,000 annotated utterances, and almost converged when trained on about 20,000 annotated utterances.

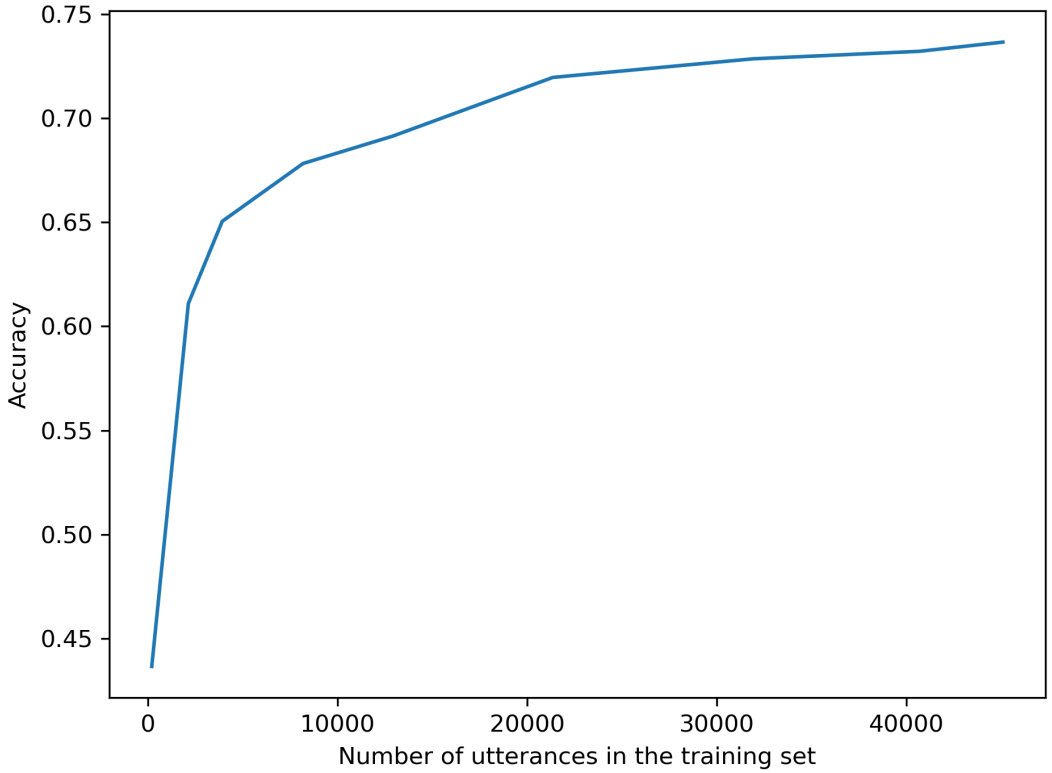


Figure 2. ***CRF: Accuracy as a function of training set size.***

Error Analysis

To gain a better understanding of our best performing model (the CRF), we perform an error analysis. For each speech act category, we calculate precision, recall and f1-score. Results can be found in the Appendix. The variance of the f1-scores for different categories is remarkably high, with values ranging from 0 to 95%. Performance is best for speech acts `QN` ("Ask a product-question") and `EA` ("Elicit onomatopoeic or animal sounds.") and worst for speech acts such as `CR` ("Criticize or point out error in nonverbal act") and `AL` ("Agree to do something for the last time.").

One important factor affecting the per-label performance is the availability of training examples and the distribution of speech acts in the dataset is heavily skewed with a long tail (see Figure 1). For labels with only very few training examples the model struggles to pick up important features. Indeed we find a high correlation between the frequency of

labels and their respective f1-score (Spearman correlation coefficient: 0.59, $p < 1 \cdot 10^{-5}$). The example in Table 2 illustrates this finding. In the conversation, all speech acts have been predicted correctly by our model except for the last utterance ("You're a nut"), which is labelled as ST ("Make a declarative statement") while the ground-truth label is DS ("Disapprove scold protest disruptive behavior"). Indeed, the speech act DS occurs very few times in the training data (only 40 examples, i.e., less than 0.1% of the training data).

Table 2: Excerpt of a conversation from the New England Corpus (Child: Liam, Age: 14 months, Transcript: 99) with manually-annotated speech acts ("Manual") and predicted speech acts ("CRF"). Labels from the INCA-A tagset are listed in the Appendix.

	Speech act	
	Manual	CRF
Mother: We're having a little problem here in the corner. (Mother stands up) (Child unplugs cord from wall again)	ST	ST
Mother: Liam ! (Mother takes hold of Child's hand)	CL	CL
Mother: No! (Mother takes hold of cord and tries to pull it out of Child's hand, Child holds onto cord)	PF	PF
Mother: Let go. (Child lets go of cord, Mother plugs cord back into wall, Child watches what Mother does with cord)	RP	RP
Mother: No. (Mother picks up Child)	PF	PF
Mother: You're a nut.	DS	ST

Another factor that affects the model's performance is what appears to be ambiguities in the definition of some categories in the INCA-A coding scheme. In particular, many pairs of speech acts are either very similar or hierarchically related (see Cameron-Faulkner and Hickey (2011) for a similar observation). More concretely, there are pairs of speech act categories that describe overlapping communicative intents (e.g., "Criticize or point out error in nonverbal act" (CR) can overlap with "Disapprove scold protest disruptive behavior" (DS) and pairs of speech acts where the meaning of one act appears to be covered by the other broader act (e.g., the speech act "Praise for motor acts i.e for nonverbal behavior." (PM) is part of "Approve of appropriate behavior." (AB)). Such overlaps in the definition of some categories do not help the model make clear distinctions between the affected categories and, thus, tend to conflate them.

We provide an example for this phenomenon in Table 3. In this conversation, the mother's utterance "Good girl" is labelled by the CRF as "Approve of appropriate be-

havior." (AB), which is not incorrect, but differs from the human annotation, which categorizes it as "Praise for motor acts i.e for nonverbal behavior." (PM). We hypothesize that collapsing overlapping categories would improve the model performance. Indeed, we experimented with an alternative coding scheme where we collapsed certain categories and the model achieves a higher average performance of 75.35% ($\pm 4.17\%$) accuracy. However, for the remainder of this work, we continue using the original coding scheme to ensure comparability to the work of Snow et al. (1996).

Table 3: Excerpt of a conversation from the New England Corpus (Child: Joanna, Age: 20 months, Transcript: 32) with manually-annotated speech acts ("Manual") and predicted speech acts ("CRF"). Labels from the INCA-A tagset are listed in the Appendix.

	Speech act	
	Manual	CRF
Mother: Take it [= book] out of the box. (*The child struggles with both hands on the open book. Afterwards, the child pulls the book up and out of the box*)	RP	RP
Mother: Good girl.	PM	AB

Replicating Findings from Snow et al. (1996)

Here we validate the CRF model by testing its ability to lead to conclusions similar to the ones obtained in Snow et al. (1996). To this end, and as we mentioned earlier, we proceed in two steps: First, we replicate major findings in Snow et al. (1996) using their hand-annotated labels. Second, we compared them to the corresponding findings obtained using the labels that were predicted using our CRF model. In addition to replicating main analyses from Snow et al. (1996) (i.e., development of the size and distribution of speech acts), we also tested the models with a new, more specific task that consists of predicting the precise normative age of acquisition of speech acts in both production and comprehension.

Development of the Number of Distinct Speech Acts

Figure 3 shows the proportion of children producing a given number of different speech act types for the three age groups studied in Snow et al. (1996) (This is a direct replication of Figure 2 in the original paper). Next to each bar obtained from the hand-annotation (in blue) we plot the corresponding bar from the automatic labeling by CRF on the same dataset (in orange).

We can see that the patterns observed in Snow et al. (1996) are well captured by automatic labeling data: At 14 months, most children produce only a handful of speech act types, such as statements (ST), repetitions (RT) and markings (MK). This number increases on

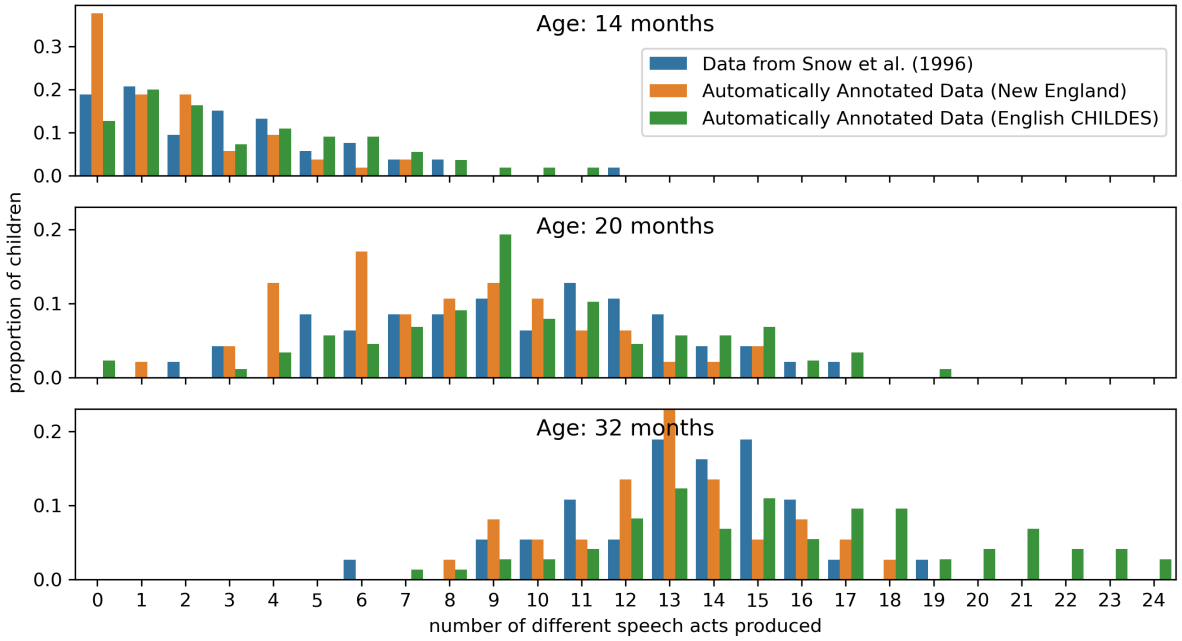


Figure 3. ***Proportion of children producing a given number of distinct speech act types at 14, 20, and 32 months old. Note that the y-axis for the bottom two figures has been shortened for better visibility.***

average for children aged 20 months where now a substantial proportion of children become able to produce around 10 different speech act types (now starting to use for example requests (RP), stating intent (ST) and product questions (QN)). Finally, at 32 months, children typically produce between 10 and 20 different speech act types (starting to use for example polar questions (YQ)). When compared to hand annotated data in the New England corpus, the model was able to capture not only the rough number of speech act types produced at each age range, it was also able to capture quite well the variability between children at each age.

We can quantify the similarity between the hand- and automatic-annotation-based distributions by computing their Jensen-Shannon distances. This measure quantifies the dissimilarity between two probability distributions with values ranging from 0 (maximally similar) to 1 (minimally similar). The similarities of distributions from manually and automatically annotated data were as follows: 0.262 (at 14 months), 0.367 (at 20 months), and 0.186 (at 32 months).

Development of the Distribution of Speech Acts

Figure 4 shows the replication of the analysis on the development of the distribution of speech acts (cf. Table 9 in Snow et al. (1996)). This analysis compares the proportions of

utterances that fall within each speech act category for the three age groups. Similar to the previous graph, next to each bar obtained from the hand-annotation (in blue) we plot the corresponding bar from the automatic labeling by CRF (in orange). We can see that the frequency distributions look remarkably similar in each age group (see Appendix for the legend of what each speech act label refers to). Jensen-Shannon distances of automatically annotated data (New England) compared to data from Snow et al. (1996) were: 0.089 (14 months), 0.103 (20 months), 0.080 (32 months).

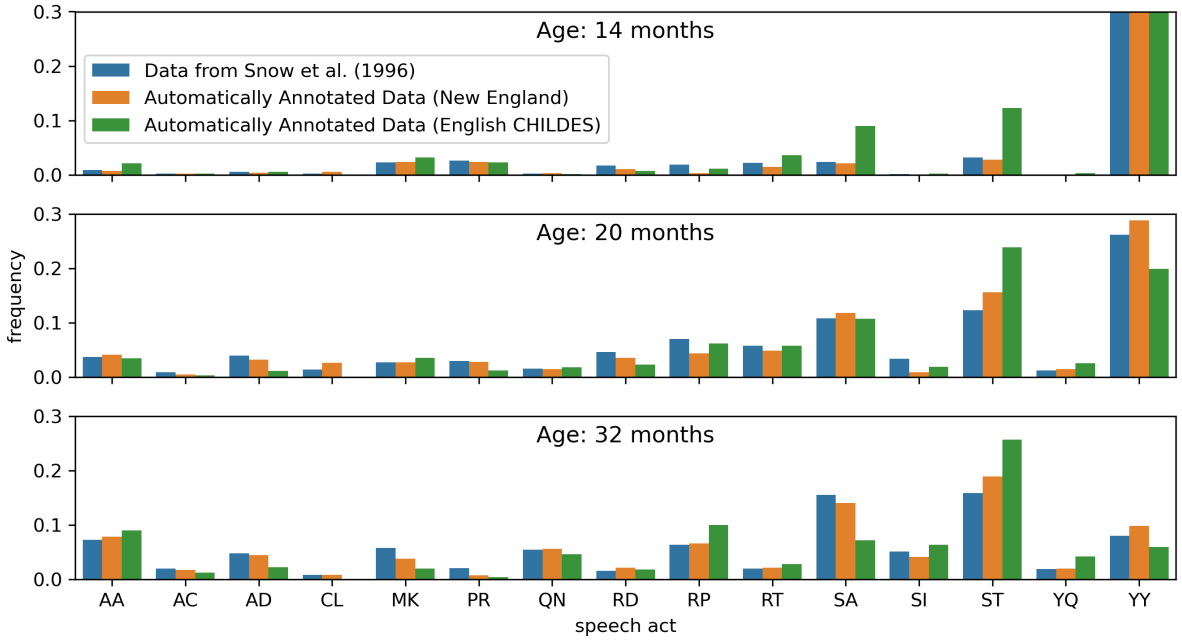


Figure 4. ***Frequency distribution of speech acts for different ages. Note that the y-axes have been trimmed for better visibility (The frequencies for*** `YY` ***at 14 months are around 0.6).***

Generalizing Findings to Data in CHILDES

In the previous subsection, we validated the model by comparing findings from predicted and hand-annotated labels of the same data. Here, we use the trained model to automatically annotate data from English corpora in CHILDES. The goal is to investigate the extent to which findings obtained in Snow et al. (1996) generalize to a larger number of children and to the variety of communicative contexts represented in these new corpora.

More precisely, we trained the CRF on the whole New England corpus (no held-out test set) and used it to annotate speech acts on transcripts of children aged between 14 to 32 months old in the North American English corpora of CHILDES (excluding transcripts

from the New England corpus). Next, we perform the same analyses as in the previous section using the large-scale annotated data.

Development of the Number of Distinct Speech Acts

The green bars in Figure 3 show the number of different speech act types produced by children from CHILDES. Developmental patterns are very similar to the original graphs (in orange), with the exception of the oldest age group (i.e., 32 months) where we found that more children produced a relatively larger number of different speech acts (more than 20). Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from Snow et al. (1996) were: 0.209 (at 14 months), 0.222 (at 20 months), and 0.418 (at 32 months).

Development of the Distribution of Speech Acts

We present the frequency distribution of speech acts for children from CHILDES in the green bars of Figure 4. Again, patterns obtained by Snow et al. (1996) generalize very well. Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from Snow et al. (1996): 0.204 (14 months), 0.173 (20 months), 0.197 (32 months).

Age of Acquisition of Speech Acts

In this section, we present results for the age of acquisition of speech acts in terms of production and comprehension using the measures defined in the Section "Measures of Speech Act Emergence".

Production

We calculated the age of acquisition for a subset of 25 speech acts[10] using both the manually-annotated labels from Snow et al. (1996) and the automatically generated labels from the CRF on the same dataset. Examples for regression plots and predicted ages of acquisition for all speech acts can be found in the appendix. Then, we calculated the Spearman rank-order correlation[11] to examine whether the *order* of emergence of speech acts is correctly captured by the automatically annotated data.

[10]These were the ones for which we could fit a logistic regression using at least two data points. While the number of acts we keep may seem small compared to the original size (65 possible speech acts excluding categories for unintelligible speech acts, `YY` and `OO`), it is due to the fact that the frequency distribution is highly skewed: Most categories occurred rarely in the corpus (Figure 1) and therefore did not provide enough data to be used in the calculation of age of acquisition.

[11]The rank-order correlation was computed over the subset of 25 speech acts for which an age of acquisition could be calculated, details in the Appendix.

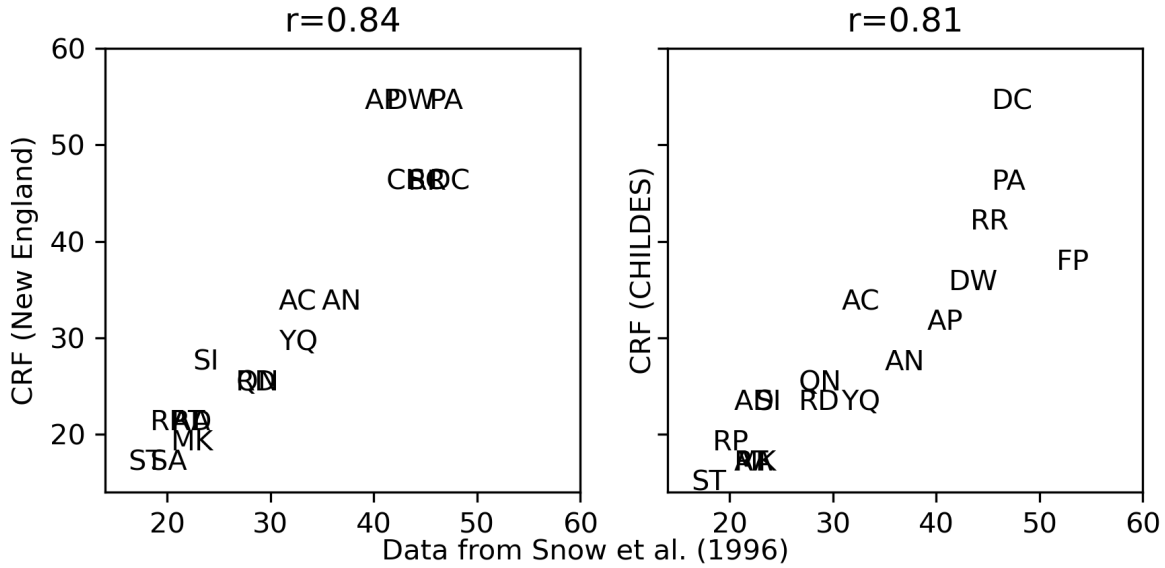


Figure 5. ***Correlation of age of acquisition in terms of production as calculated using data from Snow et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60 months for better visibility of early development. However, the correlation was calculated for all values.***

The resulting high correlation (see Figure 5 (left); $r \approx 0.84$, $p < 1 \cdot 10^{-6}$) indicates that the automatically generated labels can provide reasonable estimates for the developmental trajectory of speech acts.

We also calculated ages of acquisition using the predicted labels on CHILDES data. Figure 5 (right) shows the correlation with the ages calculated using New England data. Spearman rank-order correlation was $r \approx 0.81$ ($p < 1 \cdot 10^{-6}$).

Comprehension

To illustrate the emergence of speech acts in terms of comprehension, we first show observed adjacency pairs for adult-child turns for different ages in Figure 6. The youngest children respond with unintelligible utterances or utterances without clear function (YY, OO) in most of the cases displayed. Children at 20 months show some consistent patterns in their response behavior: Polar and product questions (YQ, QN) are answered with adequate responses (AA, SA). Polite requests (RQ) are either accepted (AD) or refused (RD). Requests or suggestions (RP) are also usually accepted or refused, although in some cases children answer with a statement (ST), which is not contingent. Additionally, there is still a large amount of utterances without clear function (YY). Only by the age of 32 months, most of the parents' utterances are addressed with contingent responses (at

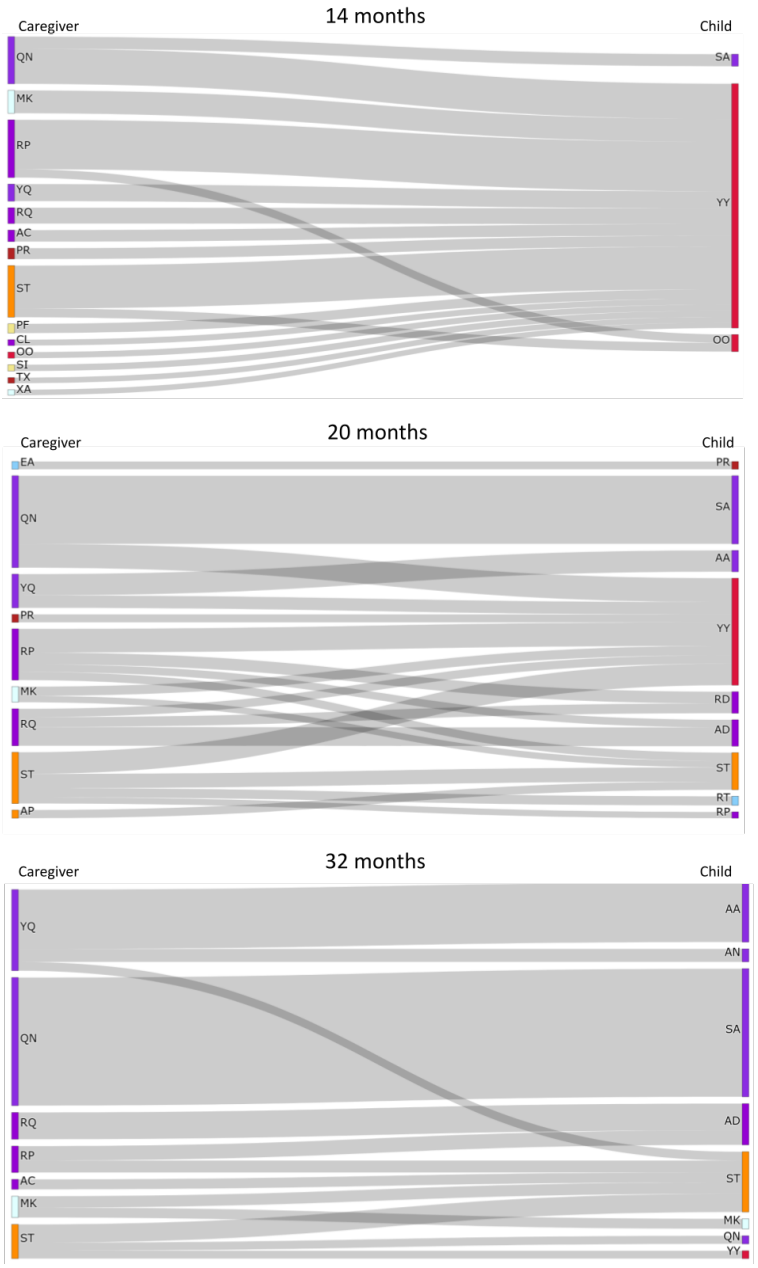


Figure 6. ***Adjacency pairs of speech acts for children of 14, 20, and 32 months. Utterances by the caregiver are on the left, responses by the children on the right. Filtered to display speech acts that occur in at least 0.01% of the data for better visibility. The colors indicate the higher-level interchange type for each speech act (see Snow et al., 1996).***

least as captured at the broad level of speech act categories).

Examples for predicted ages of acquisition for all speech acts can be found in the appendix. We observe that while there are similar trajectories in production and comprehension for some speech acts (e.g. RR), we also observed some striking differences in other cases. For example, "demands for permission" (FP) is produced very late (around 52 months), but they are already understood a lot earlier (around 14 months).

As done for the production measure, we calculated the age of acquisition using both the ground-truth labels from Snow et al. (1996) and the automatically generated labels from the CRF on the same dataset, as well as using generated labels on the English CHILDES data. As in production, the Spearman rank-order correlation coefficient[12] (see Figure 7, left; $r \approx 0.46$, $p < 0.01$) indicates a statistically significant positive correlation (however lower than for the production measure). For the correlation with predicted labels on CHILDES data, the Spearman rank-order correlation was $r \approx 0.63$ ($p < 1 \cdot 10^{-5}$; see Figure 7, right).[13]

Figure 8 shows the full distribution of age of emergence in both production and comprehension. It shows that, overall, comprehension of speech acts precedes their production. Indeed, a paired t-test (using only speech acts for which we could calculate an age of acquisition both in production and in comprehension) shows a mean difference of 2.51 months ($p < 0.05$).[14]

Finally, we ask how the trajectory of emergence in comprehension compares to that of production. For instance, does production follow the same pattern/order of comprehension, only delayed? Pearson's correlation between the two developmental trajectories is $r \approx -0.07$ ($p \approx 0.76$), indicating that speech acts emerge differently in production and comprehension, and suggesting that these two dimensions of development may be explained by different factors.

[12]The rank-order correlation was computed over the subset of 47 speech acts for which an age of acquisition in terms of comprehension could be calculated, i.e. cases in which we could fit a logistic regression using at least two data points, details in the Appendix.

[13]As we said above, we chose to fit the age of acquisition using logistic regressions following the method used for the AoA of words Frank et al. (2021). The main limitation here was the sparsity of available annotated data: The study by Snow et al. (1996) only considers 3 different age groups: Children at 14, 20, and 32 months. While the fitted curves were good for production, this was less obvious for comprehension data based on contingency (see the graphs in the appendix). Note, however, that for our analysis, i.e., correlating AoA from predicted vs. hand-annotated speech acts (Figures 6 and 7), we only needed the ranking of AoA, not necessarily absolute values of ages. So, one simple way to test the robustness of these correlations is the following: Instead of estimating the AoA using logistic regressions, we can estimate the ranking without fitting any model and directly from the data. More specifically, we computed the proportion of children that produced (or understood) a given speech act (averaged over the three-time points) and ranked the speech acts according to these proportions as a proxy for their order of acquisition. The resulting rank-order correlations obtained using this model-free method were very close to the correlations found using the regression method, thus corroborating these findings.

[14]When using the alternative coding scheme with collapsed speech act categories (see Section "Error analysis"), this difference increases to 9.61 months ($p < 0.01$).

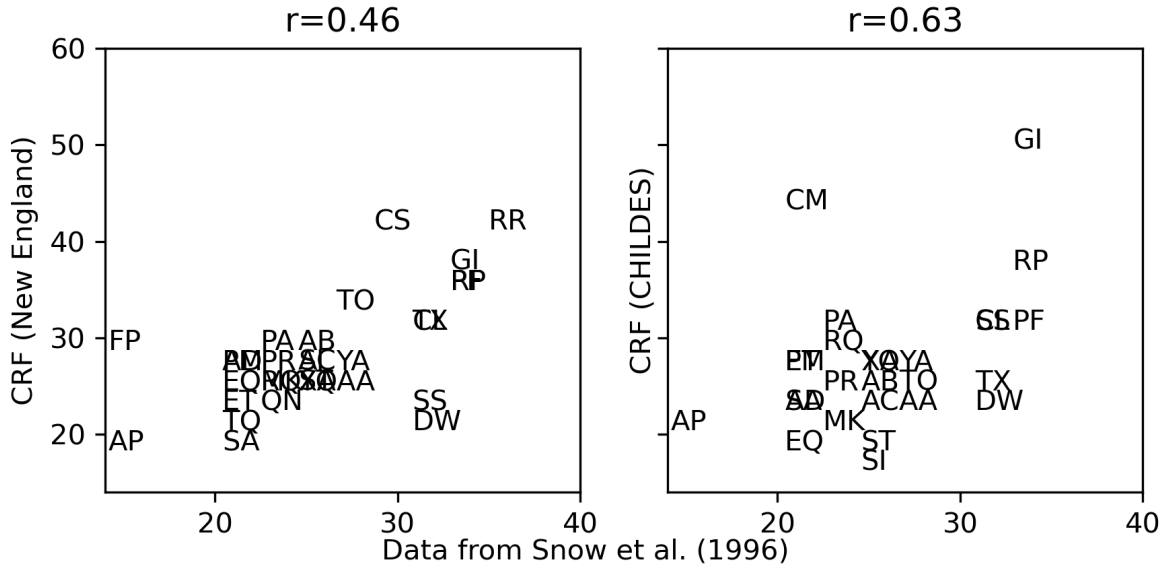


Figure 7. ***Correlation of age of acquisition in terms of comprehension as calculated using data from Snow et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60/40 months for better visibility of early development. However, the correlation was calculated for all values.***

Development of Speech Acts Beyond 32 Months

Since CHILDES contains data for children beyond the age range studied in Snow et al. (1996), we could also make predictions about the age of acquisition of some speech acts that could not be calculated using the New England corpus because they were not yet acquired by children by 32 months. To this end, we use all transcripts up to 54 months (data become sparse beyond that age). Using this larger set of annotations, we can for example estimate the age at which children produce speech acts such as prohibitions (`PF`, at 84.9 months), give reason (`GR`, at 87.0 months), polite requests (`RQ`, at 66.2 months), and make promises (`PD`, at 130.7 months)). These predictions are consistent with the developmental literature showing a late acquisition of some of these speech acts (Matthews, 2014). A table of all results can be found in the Appendix.

Discussion

The way children master language use in social interaction is an important frontier in the study of language development (Bloom & Lahey, 1978; Casillas & Hilbrink, 2020; Clark, 2018; Matthews, 2014; Snow et al., 1996). Answering this question has also the potential for impact in clinical applications (e.g., early and automatic detection of communicative difficulties). However, the investigation of this phenomenon in ecological valid settings

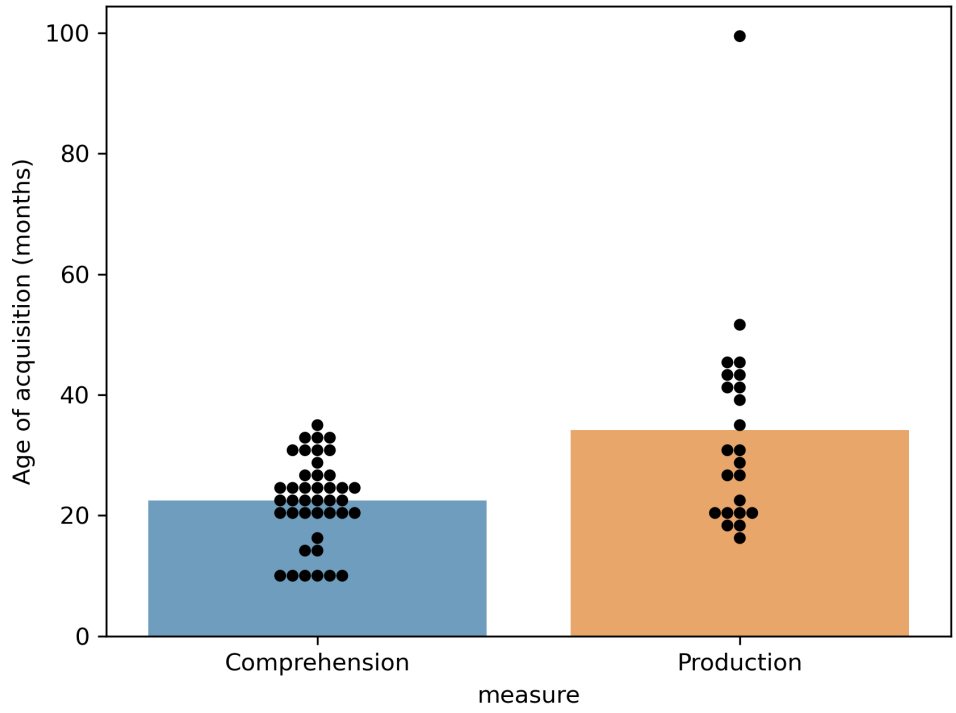


Figure 8. ***The distribution of the speech acts' age of emergence in comprehension and production.***

requires complex, large-scale data annotation which is prohibitively expensive to do by hand only.

In the current work, we introduced a simple model that allows for reliable *automatic* labeling of major speech act categories in the context of child-caregiver social interactions. We trained the model on a dataset that was previously hand-annotated using INCA-A, a comprehensive coding scheme for speech acts in early childhood (Ninio et al., 1994; Snow et al., 1996). When tested on parts of the data it had not seen in the training, the model predicted speech acts that captured quite well the major findings reported in this earlier work such as the average trajectory of speech act development and the patterns of variations between children.

Besides providing a valuable tool that we make available to the community, a major theoretical contribution of the paper was testing how earlier findings — obtained using hand annotation of a small number of children — generalize to a larger and different sample. We tested this generality by automatically labeling the entire American English section of CHILDES for speech acts. We found that, across all major analyses, children

show, overall, patterns that were very similar to the ones reported by (Snow et al., 1996). The main difference was that older children in the larger dataset produced noticeably more speech act types than children of similar age in the original study (Figure 3, bottom). This difference could be due to the fact that the larger dataset contains a richer set of conversational contexts, giving children the opportunity to perform more distinct speech act types.[15]

Anther contribution of this work is the introduction of two measures to quantify the age of emergence of speech acts in children's production and comprehension. We found that these two measures (i.e., comprehension and production) did not correlate, indicating that they provide non-redundant information about development and suggesting that speech acts may develop differently in production and comprehension. In particular, factors that would be relevant for learning in production may not necessarily be the same in comprehension, especially in the rather *asymmetrical* context of child-caregiver interactions.

To illustrate, take the case of "Yes/no requests" (`RQ`) vs. "yes/no questions for information." (`YQ`). In production, we replicated Snow et al. (1996)'s finding that children produce yes/no questions as requests later than yes/no questions for information (very few children produced the first act and only at 32 months). This fact is also in line with the literature on politeness which suggests that children produce polite requests quite late (Axia & Baroni, 1985). Interestingly however, in comprehension we found that on average children responded contingently to the yes/no requests at about the same age as they do to yes/no questions for information.

When using automatically annotated data from our model, we found that their predicted measures of age of acquisition correlated to a high degree with the ages of acquisition predicted from manually labelled data, especially in production. In a direct application, the model allowed us to estimate the age of acquisition of some late emerging speech acts (e.g., "promise" and "give reason") thanks to automatic labeling of new data children that were older in CHILDES than in the original New England corpus.

While the automatic labelling model provides a high average accuracy score, the per-label scores showed high variability. While, as we argued above, some of this variability can be explained by the frequency of occurrence in the training data and by ambiguities in the definition of some categories in the coding scheme, we speculate that other factors could be in play as well, especially the *linguistic variability* with which a speech

[15] Another observation was that the proportion of children producing no speech acts (i.e., 0 in Figure 3) at 14 months is noticeably higher in the automatically annotated data than in the original data. This means that our model classified more utterances as unintelligible or utterance without function than the human annotators. We hypothesize that the highly skewed distribution of speech acts in the dataset for children at this age, with many (but not all) utterances actually being without clear function, leads the model to overfit to this case and miss some actually meaningful utterances.

act can be expressed.[16]

For example, there is a variety of ways one can express the act of "giving reasons" (GR) in linguistic terms, which makes it relatively hard to recognize based only on the linguistic features of its instances (F-score = 0.3). In comparison, the set of linguistic terms typically used to express, say, the act of "requesting repetition" (RR) or "eliciting question" (EQ) is much more constrained, making their recognition easier (F-scores are 0.53 and 0.81, respectively), although all three categories have roughly similar (low) frequency of occurrence in the data. Take also the case of "stating intent" (SI) and "prohibiting" (PF). Both of these speech acts are similarly frequent (around 300 occurrences), but the F-score for PF is much higher than the one for SI (0.76 and 0.43, respectively). This difference could also be due to the fact that "prohibiting" is much more constrained linguistically than "stating intent."

Researchers have made a similar argument about the role that linguistic variability can have on their learnability by children (e.g. Bloom & Lahey, 1978). This analogy is to be taken with a grain of salt though. More generally, it is not warranted to make a direct link between the learnability of speech act categories by our model and their learnability by children: In the first case, the model was aimed at optimizing prediction accuracy and had been trained on labeled data. In the second case, children learn without having access to the true labels of the utterances. Models that aim at "discovering" categories in an unsupervised fashion are more likely to be insightful about the learnability of speech act categories by children (e.g. Bergey et al., 2021).

Limitations and Future Work

Our model learns how to recognize speech acts from their linguistic instances only. While the scores were quite good and allowed us to replicate major findings that were obtained using human annotations, future work should seek to build more comprehensive models that integrate multimodal cues — besides verbal language — that likely play a role in signaling communicative intents including vocal and visual cues (e.g. Fernald, 1989; Senju & Csibra, 2008; Tomasello et al., 1997; Trujillo et al., 2018). This effort will involve collecting multimodal data of spontaneous child-caregiver conversations (e.g. Bodur et al., 2021) as well as the development of machine learning methods for the automatic annotation of speech acts using linguistic, acoustic, and visual features.

Another limitation concerns the measures we used to quantify the age of acquisition. While it is easier to quantify acquisition through production, it is trickier to have a perfect measure of comprehension in a natural, uncontrolled context. Here, we provided a contingency-based measure. Such an operationalization has allowed us to uncover new

[16]Indeed, the higher the variability within a given category, the more examples the model needs to learn it.

interesting phenomena (namely that children understand some speech act before they produce them).

However, measuring contingency is a notoriously difficult task, especially in a naturalistic setting and with verbal data only. First, responses can be contingent in various ways: For example, asking a yes-no question like "Do you want a banana?" can be followed by many speech acts that can all be contingent such as "Yes!", "I just ate one", or "now?". Other speech acts such as declarative statements do not necessarily require a response, so the listener might understand the communicative intent without necessarily giving a response. In this work, we partly avoided these difficulties by using a broad binary annotation that judged whether a response was possibly contingent or totally inappropriate (e.g., a "greeting" after a "yes-no question").

In addition to these theoretical difficulties, there are practical difficulties related to the fact that children (especially the younger ones) may respond contingently but in a non-verbal fashion (a case that is not captured by the current model). Besides, they sometimes respond in an unintelligible fashion (a case which we had to classify as non-contingent). Another case is when they do not respond at all (leading to more data exclusion). However, when children do not respond (e.g., after being asked a question), it does not necessarily mean that they did not understand the speech act. For example, children may lack the appropriate vocabulary to formulate an adequate response or they may just not be interested in following up.

Finally, we did not take into account the timing of responses (as several CHILDES corpora lack timestamps in the transcripts). This is important, because if a child's response only follows a caregiver's utterance after a long temporal delay, it may not be an actually response, but a new initiation. Thus, it would not be appropriate to judge the contingency of this "response" with respect to the caregiver's utterance that preceded it.

All these reasons may contribute to making our contingency measure *under-estimate* children's early age of comprehension. That is, it is very likely that children understand many speech acts at a much earlier age than what we report in this work. That said, some results using this measure, especially the fact that comprehension precedes production in some categories, would still hold. In fact, if anything, a more accurate measure of comprehension would just make such conclusions stronger.

Finally, we found several limitations the INCA-A coding scheme when automatically labeling utterances, including overlapping as well as hierarchically related categories (cf. the error analyses section as well as Cameron-Faulkner (2014) for similar observations). In the future, the coding scheme should be updated in order to make it less ambiguous for automatic annotation.

To conclude, this work has introduced both novel research tools and measures that we hope will pave the way to a more quantitative approach to the study of children's speech act development in the wild.

References

Axia, G., & Baroni, M. R. (1985). Linguistic politeness at different age levels. *Child Development*, 918–927.

Bergey, C., Marshall, Z., DeDeo, S., & Yurovsky, D. (2021). Learning communicative acts in children's conversations: A hidden topic markov model analysis of the childes corpus. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders.*

Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). Chico: A multimodal corpus for the study of child conversation. *Proceedings of the 23rd International Workshop on Corpora and Tools for Social Skills Annotation.*

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. *CogSci*.

Cameron-Faulkner, T. (2014). The development of speech acts. *Pragmatic development in first language acquisition*, 37–52.

Cameron-Faulkner, T., & Hickey, T. (2011). Form and function in irish child directed speech. *Cognitive Linguistics*, *22*(3), 569–594.

Casillas, M., & Hilbrink, E. (2020). 3. communicative act development. *Developmental and Clinical Pragmatics, 13,* 61.

Clark, E. V. (2018). Conversation and language acquisition: A pragmatic approach. *Language Learning and Development, 14*(3), 170–185.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message?. *Child Development, 60*(6), 1497–1510.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language, 35*(3), 515–531.

Grice, H. P. (1975). Logic and conversation. *Speech acts* (pp. 41–58). Brill.

Khanpour, H., Guntakandla, N., & Nielsen, R. (2016). Dialogue act classification in domain-independent conversations using a deep recurrent neural network. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2012–2021.

Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with crf. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1).

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format.

Matthews, D. (2014). *Pragmatic development in first language acquisition* (Vol. 10). John Benjamins Publishing Company.

Ninio, A., Snow, C. E., Pan, B. A., & Rollins, P. R. (1994). Classifying communicative acts in children's interactions. *Journal of communication disorders, 27*(2), 157–187.

Okazaki, N. (2007). Crfsuite: A fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rollins, P. R. (1999). Early pragmatic accomplishments and vocabulary development in preschool children with autism. *American Journal of Speech-Language Pathology, 8*(2), 181–190.

Rollins, P. R. (2017). Pathways early intervention program for toddlers with autism. *Journal of Menatl Health and Clinical Psychology, 1*(1).

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica, 8*(4), 289–327.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, 1–23.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current biology*, *18*(9), 668–671.

Snow, C. E., Pan, B. A., Imbens-Bailey, A., & Herman, J. (1996). Learning how to say what one means: A longitudinal study of children's speech act use. *Social Development*, *5*(1), 56–84.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339–373.

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, *68*(6), 1067–80.

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A kinect study. *Cognition*, *180*, 38–51.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Data, Code and Materials Availability Statement

Source code of all models and experimentation scripts are publicly available at: https://github.com/mitjanikolaus/childes-speech-acts.

The data is publicly available as part of the CHILDES corpora. Data for the New England corpus has been directly downloaded from the CHILDES database: https://childes.talkbank.org/access/Eng-NA/NewEngland.html. Data for all other corpora has been accessed using childes-db: https://langcog.github.io/childes-db-website/.

Authorship and Contributorship Statement

M.N., E.M., J.A. and A.F. designed research; M.N. and E.M. performed research and analyzed data; and M.N., E.M., J.A., L.P., and A.F. wrote the paper.

Acknowledgements

Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-21-CE28-0005-01 (MACOMIC), AMX-19-IET-009 (Archimedes Institute) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

Appendix

INCA-A Tagset

Speech acts of the INCA-A coding scheme (Ninio et al., 1994) are listed in Table 4.

Table 4: Speech acts of the INCA-A tagset.

Speech Act	Description
AA	Answer in the affirmative to yes/no question.
AB	Approve of appropriate behavior.
AC	Answer calls/ show attentiveness to communications.
AD	Agree to carry out an act requested or proposed by other.
AL	Agree to do something for the last time.
AN	Answer in the negative to yes/no question
AP	Agree with proposition or proposal expressed by previous speaker
AQ	Aggravated question expression of disapproval by restating a question
CL	Call attention to hearer by name or by substitute exclamations
CM	Commiserate express sympathy for hearer's distress.
CN	Count.
CR	Criticize or point out error in nonverbal act.
CS	Counter-suggestion/ an indirect refusal.
CT	Correct provide correct verbal form in place of erroneous one.
CX	Complete text if so demanded.
DC	Create a new state of affairs by declaration
DP	Declare make-believe reality.
DR	Dare or challenge hearer to perform an action.
DS	Disapprove scold protest disruptive behavior.
DW	Disagree with proposition expressed by previous speaker.
EA	Elicit onomatopoeic or animal sounds.
EC	Elicit completion of word or sentence.
ED	Exclaim in disapproval.
EI	Elicit imitation of word or sentence by modelling or by explicit command
EM	Exclaim in distress pain.
EN	Express positive emotion.
EQ	Eliciting question (e.g. hmm?).
ES	Express surprise.
ET	Express enthusiasm for hearer's performance.
EX	Elicit completion of rote-learned text.
FP	Ask for permission to carry out act.
GI	Give in/ accept other's insistence or refusal.
GR	Give reason/ justify a request for an action refusal or prohibition
MK	Mark occurrence of event (thank greet apologize congratulate etc.).
NA	Intentionally nonsatisfying answer to question
ND	Disagree with a declaration.
OO	Unintelligible vocalization.
PA	Permit hearer to perform act.
PD	Promise.
PF	Prohibit/forbid/protest hearer's performance of an act

PM	Praise for motor acts i.e for nonverbal behavior.
PR	Perform verbal move in game.
QA	Answer a question with a wh-question.
QN	Ask a product-question (wh-question)
RA	Refuse to answer.
RD	Refuse to carry out an act requested or proposed by other.
RP	Request propose or suggest an action for hearer or for hearer and speaker.
RQ	Yes/no question or suggestion about hearer's wishes and intentions
RR	Request to repeat utterance.
RT	Repeat or imitate other's utterance.
SA	Answer a wh-question with a statement.
SC	Complete statement or other utterance in compliance with request.
SI	State intent to carry out act by speaker.
SS	Signal to start performing an act such as running or rolling a ball
ST	Make a declarative statement.
TA	Answer a limited-alternative question.
TD	Threaten to do.
TO	Mark transfer of object to hearer
TQ	Ask a limited-alternative yes/no question.
TX	Read or recite written text aloud.
WD	Warn of danger.
WS	Express a wish.
XA	Exhibit attentiveness to hearer.
YA	Answer a question with a yes/no question.
YD	Agree to a declaration.
YQ	Ask a yes/no question.
YY	Make a word-like utterance without clear function.

Model Details

Hyperparameters

The models were trained until convergence on a held-out dev set (10% of the training data). A small set of hyperparameter configurations based on best practices were evaluated in preliminary experiments. The configuration listed in Table 5 led to the best results.

The learning rate for training the BERT-based model is substantially lower than for the other model as this model is already pre-trained and we are only fine-tuning it on the task.

Table 5: Model hyperparameters

Hierarchical LSTM + CRF	
vocabulary size	1000
word embeddings size	200
word-level LSTM hidden layer size	200
utterance-level LSTM hidden layer size	100
dropout	0.2
optimizer	Adam
initial learning rate	0.0001
+ BERT	
same as above, except for:	
initial learning rate	0.00001

Architecture

A high-level overview of the architecture of the hierarchical LSTM+CRF model can be found in Figure 9.

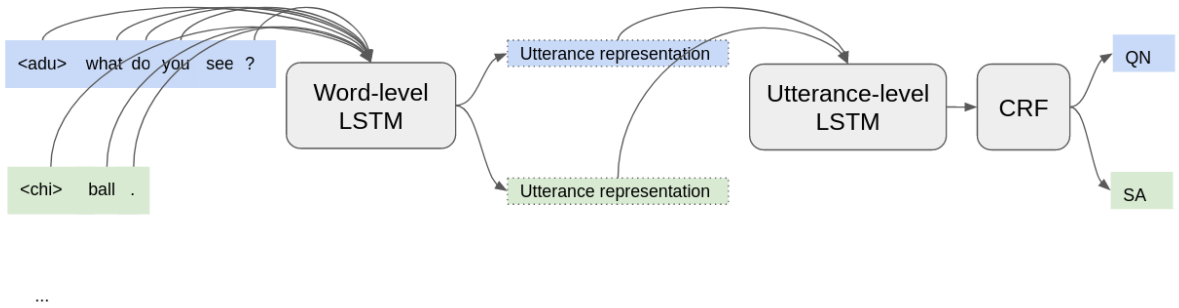


Figure 9. ***Architecture of the Hierarchical LSTM + CRF model.***

Error Analysis

Table 6 contains per-label precision, recall, and F1-scores for a model trained on 80% of the New England corpus and tested on the remaining 20%.

Table 6: Error analysis

	precision	recall	f1-score	support
AA	0.628	0.628	0.628	148
AB	0.690	0.454	0.547	108
AC	0.603	0.527	0.562	245
AD	0.674	0.651	0.662	229
AL	0.000	0.000	0.000	1
AN	0.625	0.571	0.597	35
AP	0.658	0.603	0.629	239
CL	0.800	0.875	0.836	160
CM	0.375	0.231	0.286	13
CN	0.200	0.500	0.286	4
CR	0.000	0.000	0.000	13
CS	0.273	0.086	0.130	35
CT	0.529	0.138	0.220	65
DC	0.750	0.316	0.444	19
DP	0.000	0.000	0.000	8
DS	0.375	0.273	0.316	11
DW	0.633	0.404	0.494	47
EA	0.974	0.884	0.927	43
EC	0.857	0.429	0.571	14
ED	1.000	0.333	0.500	15
EI	0.632	0.800	0.706	15
EM	0.000	0.000	0.000	1
EQ	0.750	0.849	0.796	53
ET	0.739	0.459	0.567	37
EX	0.000	0.000	0.000	1
FP	0.833	0.694	0.758	36
GI	0.375	0.158	0.222	19
GR	0.350	0.226	0.275	31
MK	0.733	0.814	0.772	996
NA	0.000	0.000	0.000	30
ND	0.000	0.000	0.000	1
PA	0.600	0.409	0.486	22
PD	0.800	0.211	0.333	19
PF	0.830	0.702	0.761	272
PM	0.518	0.345	0.414	84
PR	0.769	0.652	0.706	296
QN	0.940	0.958	0.949	1104
RD	0.679	0.494	0.571	77
RP	0.797	0.786	0.791	1689
RQ	0.830	0.848	0.839	506
RR	0.448	0.714	0.550	42

RT	0.467	0.340	0.394	144
SA	0.782	0.662	0.717	417
SC	1.000	0.455	0.625	11
SI	0.551	0.405	0.466	309
SS	0.811	0.664	0.730	116
ST	0.690	0.791	0.737	1620
TA	0.000	0.000	0.000	3
TO	0.333	0.222	0.267	72
TQ	1.000	0.200	0.333	10
TX	0.818	0.863	0.840	73
WD	0.875	0.700	0.778	10
XA	0.671	0.464	0.548	110
YA	0.769	0.408	0.533	49
YD	0.000	0.000	0.000	5
YQ	0.715	0.772	0.742	705
macro avg	0.567	0.446	0.479	10437
weighted avg	0.738	0.725	0.726	10437

Ages of Acquisition

Regression Plots

The regression plots in Figure 10 and 11 illustrate the proportion of children producing a given speech act (in the case of comprehension, the proportion of contingent responses made by children) across time as well as the best logistic fits used to predict the speech acts' precise age of acquisition. We depict only 6 exemplary speech acts for better readability. The data to create these plots was the original annotation data from Snow et al. (1996).

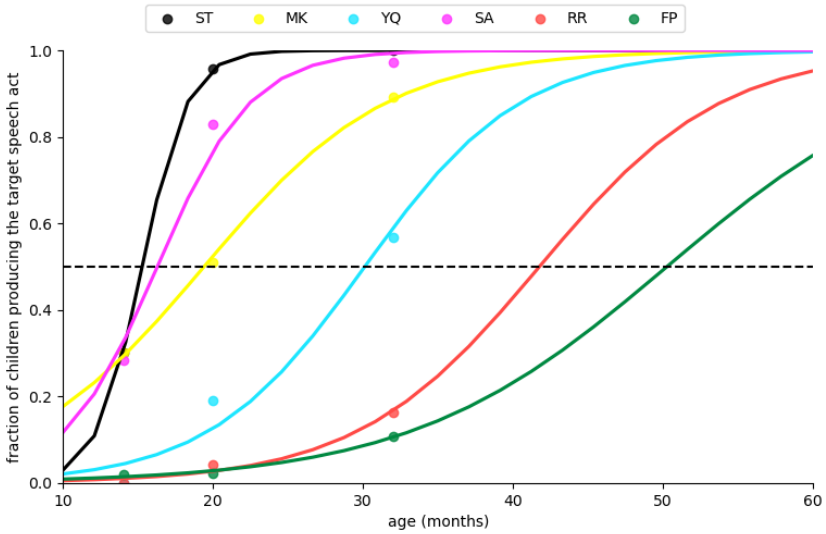


Figure 10. ***Regression plot for production.***

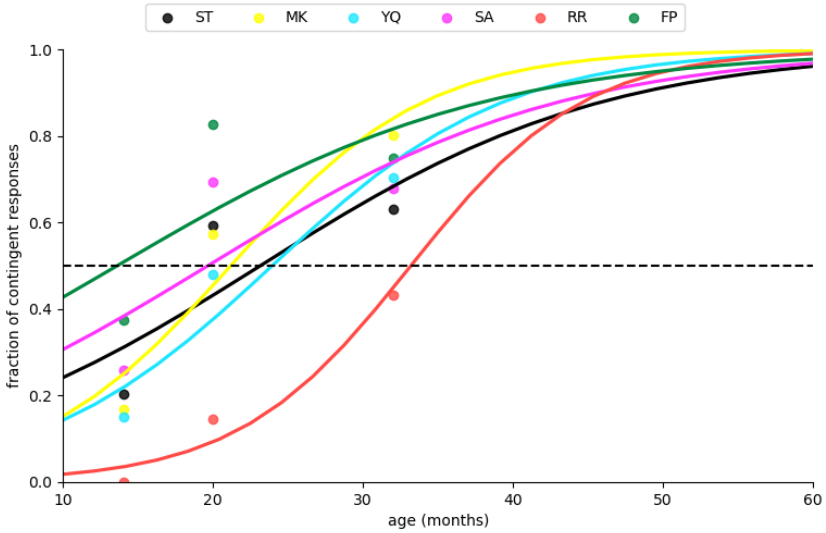


Figure 11. ***Regression plot for comprehension.***

Predicted Ages of Acquisition

The following tables show the age of acquisition (in months) for speech acts calculated using different data sources ("-" indicates that no age of acquisition could be calculated, i.e. at no observed time the proportion of children producing the speech act surpassed 0.5). We calculated the ages of acquisition in terms of production (Table 7) and comprehension (Table 8).

Table 7: Predicted ages of acquisition for production.

Speech act	Snow	CRF	CHILDES
AA	20.4	20.4	16.2
AC	30.8	32.9	32.9
AD	20.4	22.5	22.5
AN	35.0	30.8	26.6
AP	39.1	47.5	30.8
CL	41.2	45.4	70.3
CS	99.5	-	39.1
DC	45.4	45.4	53.7
DW	41.2	64.1	35.0
FP	51.6	-	37.1
MK	20.4	18.3	16.2
PA	45.4	45.4	45.4
PF	-	43.3	35.0
PR	28.7	-	-
QN	26.6	24.6	24.6
RD	26.6	24.6	22.5
RP	18.3	20.4	18.3
RR	43.3	39.1	41.2
RT	20.4	20.4	16.2
SA	18.3	16.2	10.0
SC	43.3	53.7	-
SI	22.5	26.6	22.5
ST	16.2	16.2	14.2
TO	-	35.0	37.1
YQ	30.8	28.7	22.5

Table 8: Predicted ages of acquisition for comprehension.

Speech act	Snow	CRF	CHILDES
AA	26.6	24.6	22.5
AB	24.6	35.0	24.6
AC	24.6	26.6	22.5
AD	20.4	24.6	22.5
AN	-	-	-
AP	14.2	20.4	20.4

AQ	-	-	-
CL	30.8	35.0	30.8
CM	20.4	-	43.3
CN	-	-	-
CR	-	-	-
CS	28.7	30.8	10.0
CT	16.2	16.2	10.0
DC	10.0	-	24.6
DS	-	-	-
DW	30.8	10.0	22.5
EA	10.0	12.1	10.0
EC	-	-	-
EI	10.0	22.5	10.0
EQ	20.4	22.5	18.3
ET	20.4	24.6	26.6
FP	14.2	28.7	10.0
GI	32.9	32.9	49.5
GR	10.0	26.6	20.4
MK	22.5	22.5	20.4
PA	22.5	28.7	30.8
PD	10.0	59.9	10.0
PF	32.9	26.6	30.8
PM	20.4	26.6	26.6
PR	22.5	24.6	24.6
QN	22.5	22.5	10.0
RD	10.0	-	-
RP	32.9	35.0	37.1
RQ	22.5	26.6	28.7
RR	35.0	39.1	99.5
RT	22.5	10.0	10.0
SA	20.4	18.3	22.5
SI	24.6	26.6	16.2
SS	30.8	22.5	30.8
ST	24.6	24.6	18.3
TO	26.6	35.0	24.6
TQ	20.4	12.1	10.0
TX	30.8	28.7	24.6
WD	24.6	-	87.0
XA	24.6	24.6	26.6
YA	26.6	28.7	26.6
YQ	24.6	24.6	26.6

Predicted Ages of Acquisition Including Data of Older Children

Table 9 presents the ages of acquisition in terms of production including data from older children (up to 54 months). We show only speech acts for which the age of acquisition could be calculated, i.e. for which at some age the proportion of children producing the speech act surpassed 0.5 .

Table 9: Predicted ages of acquisition including older children

Speech act	Age of acquisition
AA	18.3
AC	45.4
AD	32.9
AN	41.2
AP	101.6
AQ	155.7
CL	136.9
CN	95.3
CR	141.1
CS	149.4
DP	107.8
DW	76.6
EA	91.2
EI	164.0
EM	180.6
EQ	93.2
FP	78.7
GR	87.0
MK	16.2
PA	139.0
PD	130.7
PF	84.9
QN	35.0
RD	39.1
RP	10.0
RQ	66.2
RR	66.2
RT	10.0
SA	10.0
SI	20.4
ST	10.0
TA	95.3
TQ	62.0
YA	188.9
YQ	26.6

License

Wishes before ifs: mapping "fake" past tense to counterfactuality in wishes and conditionals

Maxime A. Tulling
Université de Montréal, Canada

Ailís Cournane
New York University, USA

Abstract: Counterfactuals express alternatives that are contrary to the actual situation. In English, counterfactuality is conveyed through conditionals ("If pigs had wings, they could fly") and *wish*-constructions ("I wish pigs had wings"), where the past tense morpheme marks non-actuality rather than past temporal orientation. This temporal mismatch seemingly complicates the already challenging task of mapping abstract counterfactual meaning onto these linguistic expressions during first language acquisition. In this paper, we investigated the role of linguistic transparency on the acquisition of different counterfactual constructions with a corpus study on the spontaneous production of English-speaking children between the ages of 2-to-6. We extracted *wish*-utterances from 52 corpora available on CHILDES to compare children's wish productions with those of adults, and additionally extracted counterfactual conditional utterances for 6 children to provide a comparative longitudinal overview of counterfactual wishes and conditionals. Our results support the idea that complexity of form-to-meaning mapping influences the emergence of counterfactual language. First, we observed a substantial number of productive errors in children's speech, where they produce counterfactuals with present tense marking instead of past. These errors are consistent with a stage where children have yet to figure out that the past tense is an obligatory component of English counterfactual constructions signaling a present non-actuality, rather than a past event on the timeline. Second, our results show that *wish*-constructions, which are linguistically more transparent than counterfactual conditionals, generally emerge before counterfactual conditionals in children's speech. This suggests that in English, counterfactual wishes might be easier to acquire than counterfactual conditionals.

Keywords: corpus, counterfactuals, form-to-meaning mapping, first language acquisition, English

Corresponding author(s): Maxime A. Tulling, Department of Linguistics and Translation, Université de Montréal, Pavillon Lionel-Groulx 3150 C9030, rue Jean-Brillant Montréal (QC) H3T 1N8, Canada. Email: maxime.tulling@umontreal.ca

ORCID ID(s): https://orcid.org/0000-0001-7655-9095; https://orcid.org/0000-0002-4288-4049

Citation: Tulling, M.A., & Cournane, A. (2022). Wishes before ifs: mapping "fake" past tense to counterfactuality in wishes and conditionals. *Language Development Research*, *2*(1), 306—355. https://doi.org/10.34842/2022.0559.

Introduction

Counterfactual reasoning encompasses our ability to think about alternative ways the world could be or could have been. With counterfactual expressions such as "If pigs had wings, they could fly" or "I wish pigs could fly" we express situations that are contrary to the actual state of affairs (pigs do not have wings) and imagine what the world would look like if they were true. In language development, the acquisition of counterfactuality is dependent on both cognitive and linguistic development. On one hand, children need to acquire the ability to postulate the non-actual alternative in conjunction with the actual state of affairs, which is typically thought to be a cognitively demanding task (Beck et al., 2009; Byrne, 2007). On the other hand, children need to acquire the linguistic structures that express counterfactuality in their language, and map counterfactual meaning onto these linguistic expressions. While various studies have investigated the acquisition of counterfactuality in production (e.g., Bowerman, 1986; Kuczaj & Daly, 1979; Reilly, 1982) and comprehension (e.g., Nyhout & Ganea, 2019; Rafetseder et al., 2010; Riggs et al., 1998; Robinson & Beck, 2000), we know little about the interaction of cognitive complexity, linguistic complexity, and form-to-meaning mapping in children's development of counterfactual reasoning.

The complexity of the form-to-meaning mapping of abstract concepts, is often thought to be dependent on input availability and the transparency of the linguistic cues that signal abstract meaning (Slobin, 1973, p. 178; Weist et al., 1997). Linguistic constructions that are transparent or dedicated in their expression of a complex concept are thought to facilitate language acquisition. In this paper, we explore this hypothesis by investigating the emergence of counterfactual language in the spontaneous production of English-speaking children between the ages 2-to-6. Specifically, we consider the influence of potentially misleading cues (the counterfactual's "fake" past tense) and the role of construction transparency (whether an expression is dedicated to expressing counterfactuality or not) on the acquisition of counterfactual constructions. Before we get more into the details of our study, we will first discuss the definition of counterfactuality and counterfactual reasoning, and provide background on children's acquisition of counterfactuality.

Defining Counterfactuality, Counterfactual Reasoning, Imagination and Desire

The Expression of Counterfactuality

Counterfactuality is a grammatical category used for linguistic expressions that imagine situations that are contrary to fact and different from the current or past situation (Iatridou, 2000). In English, counterfactuality can be expressed through

counterfactual (CF) conditionals (1) and wishes (2), making reference to an alternative present (1a/2a) or past (1b/2b). Crucially, the utterances in (1a-2b) all discuss an imagined car possession, while implicitly asserting that the speaker did not own a car at the reference time.

(1a) If I **had** a car right now, I would drive. PRESENT CF CONDITIONAL
(1b) If I **had** had a car back then, I would have driven. PAST CF CONDITIONAL

(2a) I wish I **had** a car right now. PRESENT CF WISH
(2b) I wish I **had** had a car back then. PAST CF WISH

Closely related to the present and past counterfactual, there is the future "counterfactual" or 'future less vivid' (FLV) (Iatridou, 2000). This construction (3) can strictly not be called counterfactual, as it refers to the future and is in principle still realizable[1]. In counterfactual conditionals (3a), the future reading is the result of the eventive main verb in the *if*-clause (e.g., *went)*. In wishes, the future reading comes from the inclusion of the verb *would* (3b). Like the present and past counterfactual, the future less vivid indicates the speaker believes the opposite to be most likely true (e.g., the utterances in (3) can be used when someone is scheduled to leave next week instead).

(3a) If he **went** tomorrow, he would get there next week. FUTURE LESS VIVID (FLV)
(3b) I wish he **would** go tomorrow. (Iatridou, 2000, 28)

The counterfactual and FLV utterances above, have in common that they all include past tense marking (indicated in bold). Usually, past tense inflection indicates an actual past, and can only combine with a temporal adverb that matches this temporal orientation, like *yesterday* (4).

(4) I **had** a car (*right now/*tomorrow/yesterday).

[1]However, Iatridou (2000, p.235) raises the question of whether we should be considering it a real counterfactual after all, as it patterns alike with the other constructions. In wishes, future temporal orientation seems to indicate a desire to change a future that the speaker believes to be unlikely or impossible to change, e.g., because it's planned or determined.

However, in counterfactual constructions the past morpheme gives rise to a non-actual interpretation instead (Iatridou, 2000; Ippolito, 2006; Karawani & Zeijlstra, 2013; Ogihara, 2000; Romero, 2014). This mismatch between the counterfactual's morphological tense marking (past) and temporal orientation (dubbed "fake" past tense by Iatridou, 2000), becomes evident when the "fake" past is combined with the present temporal adverb *right now* (1a/2a) or future temporal adverb *tomorrow* (3). In order to express true past temporal orientation (1b/2b), counterfactuals require double past marking (both "fake" and actual past) in the form of the 'past perfect'.

The occurrence of a "fake" past tense in counterfactual utterances is fairly prevalent across distinct language families (Bjorkman & Halpert, 2017; Iatridou, 2000; James, 1982; von Prince, 2017, p.6 and references therein). For this reason, it is often theorized that the "fake" past plays an important function in the linguistic expression of counterfactuality. There are two main approaches to analyzing the semantic role of the counterfactual's past tense morpheme. Past-as-past (or 'back-shifting') approaches argue that the counterfactual's past tense morpheme fulfills the function of shifting back in time (e.g., Dudman, 1983; Ippolito, 2006; Ippolito & Keyser, 2013; Ogihara, 2000; Romero, 2014), while past-as-modal ('remoteness-based') approaches believe the counterfactual's past is "fake" in the sense that the morpheme does not make any temporal reference (Bjorkman & Halpert, 2017; Iatridou, 2000; Karawani, 2014; Karawani & Zeijlstra, 2013; Ritter & Wiltschko, 2014; Schulz, 2014). For example, Iatridou (2000) argues that the past tense morpheme is the realization of an 'exclusion' feature, that either scopes over times (excluding the present, resulting in a past tense reading) or over worlds (excluding worlds, resulting in a counterfactual reading). For our purposes, we are not committed to a specific semantic analysis. Instead, we hope to have illustrated that the expression of counterfactuality is a linguistically complex phenomenon, that requires figuring out the non-transparent mapping of counterfactuality to a morpheme that usually expresses past temporal orientation and learning the semantic operations supporting this counterfactual interpretation.

Counterfactual Reasoning

Besides the linguistic complexity of expressing counterfactuality, the cognitive processes underlying counterfactual thought are complex as well. Counterfactual reasoning refers to the cognitive ability to imagine counterfactual situations. In a narrow sense, this only includes thoughts about "what might have been", which are thoughts about alternatives to specific elements of the actual world (Beck, 2016). Such counterfactual reasoning is thought to involve the ability to hold multiple possibilities in mind, while temporarily considering a false possibility as true (Beck et al., 2009; Byrne, 2007). While the linguistic concept of counterfactuality includes both the

imagination of alternative states in the present and past, developmental psychologists often define counterfactual reasoning more strictly as 'undoing a past event, action or state', requiring the consideration of two alternative representations of the same past time (Byrne, 2007; Rafetseder et al., 2010; Rafetseder & Perner, 2012; Robinson & Beck, 2000). However, it is important to note that counterfactuals expressing alternative present states (1a/2a) involve the same core processes of counterfactual reasoning, namely keeping in mind two conflicting representations and temporarily undoing what is known to be true about the actual state. For this reason, we will use the term 'counterfactual reasoning' to include the undoing of actions, states and events in the past, as well as the undoing of present states. By including present counterfactuality in the consideration of the development of counterfactual reasoning, we can isolate the mental operation of counterfactual reasoning. That said, past counterfactuality is arguably more cognitively demanding than just reasoning counterfactually about the now, because it requires the child to combine the mental operation of counterfactual reasoning with mental time travelling.

Pretend Play and Counterfactual Reasoning: Where to Draw the Line?

Besides the narrow definition of counterfactuality discussed above, some researchers use the term 'counterfactual reasoning' to include all types of 'unreal' thinking, including pretense, future thinking and reasoning about fictional worlds, as well as counterfactual reasoning in the narrow sense (Beck, 2016). Specifically, pretend play and counterfactual reasoning are thought to rely on the same cognitive abilities. Both types of thinking involve disengaging with current reality, postulating and reasoning about an alternative reality, and keeping the alternative possibility separate from reality (Walker & Gopnik, 2013; Weisberg & Gopnik, 2016). For this reason, it has been suggested that pretend play might be an important precursor to imagining possible worlds (Francis & Gibson, 2021; Gopnik & Walker, 2013). Supporting this view, some studies have found a correlation between children's performance on reasoning tasks that involve pretending and tasks that involve counterfactual reasoning (Buchsbaum et al., 2012; Francis & Gibson, 2021). In fact, Walker and Gopnik (2013) argue that pretending is a form of counterfactual reasoning, and that pretend play provides early opportunities to learn and develop this skill. However, this inclusion of pretense into the definition of counterfactuality seems to be too generous. Beck (2016) argues that pretend play and counterfactual reasoning are quantitively different in their relationship with reality and the cognitive demands they make. Beck (2016) points out that real-world counterfactuals are closely tied to reality while pretend play is decoupled from reality, and therefore does not make the same cognitive demands. In other words, pretend play is achieved by temporarily shifting into an alternative here-and-now, while counterfactual reasoning requires the postulation and comparison of

possible worlds incompatible with the actual one (Tulling, 2022, p. 175). In this paper we therefore use the definition of counterfactuality as discussed above.

The Difference between Wishing and Desiring

As discussed earlier, counterfactuality can be expressed in English using both counterfactual conditionals and wishes. While counterfactual conditionals involve causal reasoning ("If…then…"), counterfactual wishes involve the expression of desire. In English, counterfactual desire is expressed by the verb *wish* embedding a finite sentence, representing a full proposition, e.g., "I wish [I had a dog]". Note that while the verb *wish* sometimes occurs with other complements, like a Noun Phrases (NP) ("I wish you a happy birthday"), Verb Phrases (VP) ("I wish to sleep") or Prepositional Phrases (PP) ("I wish for more presents"), these uses are not counterfactual and are structurally distinguishable from propositional embedding *wish* (Iatridou, 2000, p. 241). Not all languages have a word that specializes in expressing counterfactual desire, and languages like Dutch or Greek for example use the regular desire verb *want* for this purpose. In English, both *wish* and *want* express desire, and occur with multiple different complement types, however they are distinct in both their structure and their meaning. Propositional embedding *wish* selects for a counterfactual complement and can only express desires that are non-actual and thought to be out of reach. The verb *want* selects for verbal complements with a future orientation. The desire expressed by *want* may or may not be fulfilled in the future, and can be either achievable (e.g., "I want to eat an apple") or impossible (e.g., "I want to grow wings"). The counterfactual component of the propositional *wish*-construction in contrast to a regular desire becomes obvious when we try to combine desires with their outcomes. You can *want* things you already have (5a), but it is impossible to wish for things you already have (5b).

(5a) I live in Bolivia because I want to live in Bolivia. (Iatridou, 2000, 38)
(5b) *I live in Bolivia because I wish I lived in Bolivia. (Iatridou, 2000, 40)

Acquiring the counterfactual *wish*-construction thus requires the child to learn that the verb *wish* differs from desire verbs like *want* in its counterfactual implication and can only be used when the desire is believed to be unfulfilled. We discuss the challenges to mapping counterfactuality onto linguistic expressions in more detail later. Before this, now that we have all relevant definitions in place, we provide an overview on prior research on children's acquisition of counterfactuality.

Background: Children's Development of Counterfactuality

Prior research shows that children generally start producing and comprehending counterfactual conditionals around age 4, after they have developed the ability to refer to hypothetical future events (such as "If it rains tomorrow, we will play inside"), which already seem to be in place by age 3 (Bowerman, 1986; Guajardo et al., 2009; Nyhout & Ganea, 2019; Reilly, 1982; Riggs et al., 1998; Robinson & Beck, 2000). This asymmetry between the acquisition onset of the hypothetical future and counterfactual has mainly been attributed to the additional cognitive load demanded by counterfactual reasoning, which depends not only on holding multiple possibilities in mind, but also requires temporarily considering a false possibility as true (Beck et al., 2009; Byrne, 2007). Reilly (1982) used longitudinal recordings and diary entries about one child (Kate), and various elicitation tasks with children between age 2-to-8 to investigate their acquisition of conditionals. She found that most children produce hypothetical conditionals by age 3 but did not yet fully comprehend hypothetical conditionals by this age and did not seem to understand counterfactuals. In fact, when asked counterfactual "what if" questions, many 2-year-olds and quite some 3-year-olds denied the counterfactuals or responded to them as if they were about reality, see (6a) and (6b):

(6a) Adult: What if a snake had eaten your daddy? (Reilly, 1982, ex. 37 p.107)
Cate (2;8): No! / Can't eat my daddy

(6b) Adult: What if you were a snake? (Reilly, 1982, ex. 57 p.116)
Janine (3;0): I'm not a snake / I'm Janine.

At age 4, Reilly (1982) found that children demonstrated comprehension of both hypotheticals and counterfactuals. They no longer denied the possibility of a situation or gave realist replies to counterfactual utterances. They also produced clear spontaneous present counterfactual conditionals (7).

(7) 4-yo: If they put a goldfish in there and they ate it, they would die.
(Reilly, 1982, ex. 68, p.121)

Kuczaj & Daly (1979) investigated the longitudinal development of Abe and did a cross-sectional study of 14 other children. They similarly found that future hypothetical conditionals seem to be acquired before counterfactual conditionals and reported that Abe used his first past counterfactual conditional at the end of age 3 (3;11).

The age at which children start producing counterfactual conditionals thus seems to align with when they are found to start understanding these constructions, around age 4. However, in a corpus study of three children, Bowerman (1986) noted some surprising instances of (present) counterfactual conditionals at age 2 (8a,b), and also noticed children using counterfactual *wish* at this age as well (9).

(8a) <Just having crossed a narrow street when a car goes by> (Bowerman, 1986, 43)
Christy (2;4): That car [will/would?] hit me if I was in a street

(8b) <Child is tired during long wait in doctor's office> (Bowerman, 1986, 44)
Eve (2;11): If we (didn't?) have to wait for so long
we would have be gone a long time

(9) Christy (2;1): I *wish* Christy have a car (Bowerman, 1986, 10)
I *wish* me have a airplane

While prior corpus studies mostly focused on the acquisition of past counterfactual conditionals, simpler counterfactual constructions such as the present counterfactual conditional (lacking the past perfect) or counterfactual *wish*-construction (dedicated counterfactual construction) might thus be available to children at an earlier age. This would be in correspondence with findings about spontaneous modal productions, where the linguistically less complex modal adverbs were found to be acquired before modal auxiliaries for inferential meanings (Cournane, 2021). Notably, the *wish*-utterances in (9) lack the obligatory "fake" past tense and use the present tense verb 'have' instead. This suggests that the "fake" past is a complex feature of counterfactuality, one that children initially may struggle with. In the next section, we discuss how the linguistic complexity of the "fake" past and the transparency of different constructions may influence the acquisition of counterfactual constructions.

Mapping Challenge: Attributing Counterfactual Meaning to the "Fake" Past Tense

Besides developing the cognitive mechanisms and conceptual structures necessary to support counterfactual reasoning, the acquisition of counterfactuality also requires mapping counterfactual meaning onto linguistic expressions. Children have to derive from their input which structures in their language(s) express counterfactual meaning and acquire the linguistic mechanisms that support the expression of counterfactuality (Clark, 1987; Slobin, 1973; Weist, 2018). As for this form-to-meaning mapping, there are three properties of counterfactual constructions that make this mapping particularly challenging. First, it is not obvious how children learn to map meaning onto linguistic forms when the expressed meaning is not perceptually observable

(Gleitman et al., 2005; Landau & Gleitman, 1985). In the case of counterfactual constructions (e.g., "I wish we had a dog"), this is particularly true, as by definition the proposition expressed by the counterfactual is not true in the actual world, and thus cannot be observed. Second, there is no one-to-one correspondence between form and counterfactual meaning (Clark, 1987). Counterfactuality can be mapped onto different types of linguistic expressions, such as counterfactual conditionals or wishes and also involves attributing more than one abstract meaning, past temporal orientation and the counterfactual "fake" past, to the same morpheme. Third, the counterfactual meaning of the past tense morpheme is less common, and more restricted in its environment than the regular past temporal orientation meaning. In their acquisition of counterfactuality, children thus have to learn in exactly what contexts the past tense morpheme, which predominantly expresses past temporal orientation, is "fake" and fulfills a counterfactual function instead. How do children figure this out?

Recurrent exposure to counterfactual situations described by counterfactual utterances should allow a child to pick up on the linguistic devices used to express counterfactuality. If a construction is dedicated to express counterfactual meaning, in other words it only expresses counterfactuality, it should be easier to detect from the input and link to the counterfactual situation than expressions that are used in a wider range of situations. In English, it therefore seems that counterfactual wishes should be easier to detect than counterfactual conditionals. As discussed before, the *wish*-construction is a dedicated construction in English. Whenever the verb *wish* embeds a propositional complement, this proposition is interpreted counterfactually (10a). Because of the *wish*-construction's dedication to counterfactuality, which requires usage of the "fake" past, *wish* cannot co-occur with a present tense complement in standard varieties of English (10b). This is in contrast with conditionals, where the complementizer *if* can introduce both hypothetical conditionals (11a/b) and counterfactual conditionals (11c) and co-occurs with both present and past inflected verbs.

(10a) I wish I **had** a car.
(10b) *I wish I **have** a car. (Iatridou, 2000, 25)

(11a) If he **has** time to bake cookies, he will bring some. PRES. CONDITIONAL
(11b) If he **had** time to bake cookies, he will bring some. PAST CONDITIONAL
(11c) If he **had** time to bake cookies, he would make some. PRES. COUNTERFACTUAL

The consistent usage of the "fake" past tense in the *wish*-clause, even when there is a salient mismatch between the temporal orientation and morphological past marking of the *wish*-complement (10b), may cue the child to realize its role in expressing counterfactual meaning. Conditionals that can appear with present (11a), real past (11b)

and "fake" past tense (11c) in their antecedent, render the input less transparent to discover that the counterfactual conditional's past tense does not simply indicate a true past temporal orientation. In order to know the past in (11c) is "fake", one has to link the first clause with the second containing *would*, which requires keeping in mind and causally relating two clauses (c.f. Reilly, 1982; Bowerman, 1986). The *wish*-construction lacks such causal dependency. Combined with the fact that proposition-embedding *wish* is a dedicated counterfactual marker and consistently appears with the "fake" past in the child's input, the form-to-meaning mapping of this construction can be considered less complex than that of counterfactual conditional constructions.

Aims and Hypotheses

As we have seen so far, children appear to acquire counterfactual past conditionals relatively late compared to future hypothetical constructions (e.g., Bowerman, 1986; Reilly, 1982; Riggs et al., 1998; Robinson & Beck, 2000). What makes counterfactuals more complex? In order to acquire an abstract linguistic construction involving complex reasoning, two criteria need to be fulfilled: 1) the child must have developed the cognitive ability to support the mental operations involved in representing the meaning of the utterance, and 2) the child must figure out which linguistic forms are used to express such meanings in their target language(s) (Clark, 2001; Reilly, 1982). As for the cognitive factors underlying counterfactual reasoning, an immature development of executive functions like working memory, attention switching and inhibition have been linked to the late acquisition of counterfactuality (Beck et al., 2009; Beck, Riggs, et al., 2011, p. 20; Byrne, 2007; Guajardo et al., 2009; Robinson & Beck, 2000). A cognitive leap around the age of 4 would allow children to start reason counterfactually. While this generally aligns with the age children have been found to start producing past counterfactual conditionals (Bowerman, 1986; Kuczaj & Daly, 1979; Reilly, 1982), there have been some examples of children using simpler present counterfactual conditionals and wishes at age 2, sometimes lacking the "fake" past tense (Bowerman, 1986). However, it is not certain whether these findings are exceptional, or part of a more widespread pattern in development. In this paper, we investigate the emergence of counterfactual language with a corpus study on the spontaneous production of English-speaking children between the ages of 2-to-6. Specifically, we aim to explore the role of form-to-meaning mapping in the acquisition of counterfactuality by investigating linguistic transparency from two angles.

First, we investigate the role of the counterfactual's "fake" past tense in the acquisition of counterfactual constructions. In English, counterfactual utterances contain past tense marking, even if the utterance is about the present. The fact that counterfactuality maps to the same morpheme as past temporal orientation is not only

opaque, but also potentially misleading as children might initially hypothesize that the counterfactual's past tense marking indicates past tense meaning. When do children realize the counterfactual past expresses counterfactuality rather than past temporal orientation and is a necessary component of counterfactual utterances? To investigate this question, we will examine children's spontaneous productions of counterfactual wishes. Unlike conditionals, *wish*-constructions in standard varieties of English cannot take on present tense in their complements. This means that the child's input will always contain utterances such as "I wish I *had* a dog" and not "*I wish I *have* a dog". If children mimic their input, or immediately realize the past tense morpheme belongs to the expression of counterfactuality, we expect children to match their input in their own productions. That is, when expressing a desire about the present, they will use the *wish* + "fake" past construction. However, if children go through a stage where the mapping between the "fake" past tense and counterfactuality is not yet clear, they might initially mistake the counterfactual's past in their input as referring to past situations. In this scenario, their underlying representation of the *wish*-construction would not include the "fake" past as an obligatory component, and we expect that they would mark their own spontaneous wishes just like they would in other contexts: using past tense to express desires about the past and using present tense to express desires about the present. We therefore predict children to produce non-adultlike utterances, such as "I wish I *have* a dog".

If they do, then a secondary question is whether their non-adultlike constructions are used in adultlike counterfactual contexts. Is realizing the counterfactual function of the past morpheme a necessary prerequisite for expressing counterfactuality? If it is, tense errors are expected to indicate a non-adult like use of the counterfactual *wish*-construction. For example, if a child produces "I wish I have a dog", this use of *wish* with a present-marked or bare verb complement could indicate a simple desire, in line with non-counterfactual desire verbs like *want* or *hope*. Alternatively, it could be that the "fake" past is not a necessary component of the *wish*-construction, and that children map counterfactuality only to the word *wish* inside this construction. In this case, we expect that non-adult like utterances such as "I wish I have a dog" can be used in adult-like counterfactual contexts. To find an answer to these two questions, we extract all children's *wish* usages and code for present-for-past tense errors as well as the linguistic and situational context of counterfactual usage. To gain more insight into the overall properties of children's *wish*-productions, we also compare their productions against the adult input and provide an overview of various semantic and syntactic variables.

Second, we investigate the role transparency and dedication to counterfactuality plays in the acquisition of counterfactuality. It is generally thought, that linguistic

expressions that are dedicated to expressing some type of complex abstract meaning are easier to acquire than more opaque constructions expressing that same meaning with more complex form-to-meaning mapping (Rett & Hyams, 2014; Slobin, 1973; Weist et al., 1997). As laid out in the previous section, in English, wishes are dedicated counterfactual constructions, while conditionals are not. Does this then mean that counterfactual wishes are easier to acquire than counterfactual conditionals? If it does, we expect children to start producing counterfactual wishes before counterfactual conditionals, as the form-to-meaning mapping task for this construction is more straightforward and transparent. Such a finding would indicate that it is not just conceptual development that determines the onset of counterfactual constructions in children's productions, but that linguistic factors influence the onset of different constructions. If on the other hand, children start producing both constructions around the same time, or produce counterfactual conditionals before their wish counterparts, it suggests that linguistic transparency does not play as big of a role in the acquisition of these counterfactual constructions, and that any onset differences may be the result of other cognitive factors at play. In order to address this question, we look at the longitudinal counterfactual development of six children and compare the onset of counterfactual conditionals and *wish*-constructions.

Methodology

Part 1: Children's and Adult's Wishes and the "Fake" Past Tense

Selection Criteria & Preprocessing

We looked at natural child productions of counterfactual constructions by searching through English corpora of transcribed children's speech available on CHILDES (MacWhinney, 2000) using the database 'childes-db' (Sanchez et al., 2019), accessed through the statistical software environment R (R Core Team, 2021). All operations involving corpus extraction were performed using the analysis package 'childesr' (db version = "2020.1"). We selected corpora that contained data from typically developing monolingual children between 2;5-6;0, yielding 57 corpora (48 from Northern America, 11 from the United Kingdom) including data from 585 children in total. In Appendix S1 you can find an overview of all corpora used.

For these corpora, we extracted all utterances and calculated the amount of child and adult utterances. For this calculation, speakers with the speaker roles "Target Child", "Child", "Sister", "Brother", "Friend", "Playmate", "Girl" and "Sibling" were included in the child category, while all other roles we treated as adults. We noticed that a small proportion of the data (77551 utterances, 3.5%) across 15 different corpora (partially)

lacked age information for the children in the output of the 'get_utterances( )' function. Most missing age data (2.5%) could be recovered from a participant overview extracted with the function 'get_participants( )', and for the remaining 13 corpora that still (partially) lacked target child age information we manually recovered the information where available by retrieving it from the CHILDES Talkbank corpus description pages on https://childes.talkbank.org/access/. For two corpora (MacWhinney and Gathercole) age information was displayed incorrectly (based on the metadata available in the corpus descriptions), so this was manually corrected by extracting the info from the corpus description pages (Gathercole) or recalculating the children's ages based on the transcript file name (which was based on the age of the child 'Ross', so in order to calculate the age of his younger sibling 'Mark' we subtracted 01;10;25). We then filtered the data set to only include utterances from children who were within our age-range of interest 2;0-6;0 and proceeded to extract all child utterances containing the word *wish*. In total, 40 of the searched corpora contained child wishes. For these 40 corpora we also extracted all adult utterances (child-directed speech and speech addressed to other adults within the child's hearing), so we could compare *wish* usage between children and adults.

Exclusions

To get an idea of the proportion of *wishes* present in spoken child and child-directed speech, we calculated the percentage of *wish* utterances for the child and adult corpora. We extracted 478 child utterances containing *wish* (0.02% of 2,247,665 total utterances) coming from 40 different corpora, and 841 adult *wish*-utterances (0.03% of 2,934,114 total utterances). To make a fair comparison between the *wish*-productions of children and their input (child-directed or overheard adult *wish*-utterances), we only analyzed adult data from the 40 corpora we found child wishes in. For the adult utterances, we thus proceeded to exclude 70 utterances that came from corpora that did not yield any child wishes. For the child utterances, we excluded 10 child wishes for which the target child's age was unknown. For the remaining 468 child and 771 adult *wish*-utterances, we first excluded all utterances in which *wish* was used as a noun (e.g., "Do you want to make a wish?"), which resulted in 29 exclusions for child utterances and 129 for adults. Since the verb *wish* is counterfactual only if its complement is a full proposition (Iatridou, 2000, p.241), we then excluded utterances where *wish* did not embed a proposition. For children, this resulted in 58 exclusions (2 VP complements, e.g., "not wish to play"; 17 NP complements, e.g., "I wish you a happy birthday"; 5 PP complements, e.g., "I wish for daddy to come home" and 34 instances where there was no complement, e.g., "yeah I wish"). For adults we excluded 142 non-propositional complements (11 VP, 69 NP, 13 PP and 49 missing embeddings). Lastly, we excluded an additional 32 child wishes and 15 adult wishes for being a repetition

of either themselves or someone else. This means that in total 349 child wishes and 485 adult wishes remained for further analysis.

Coding Conventions

All *wish*-utterances were manually coded for various structural and semantic linguistic variables. Structural linguistic variables included: person of the main subject, i.e., 'the wisher' (*I* and *we* = 1st person; *you* = 2nd person; *Mommy, he* and *the cat* etc. = 3rd person; no subject = omitted; inaudible subjects = unclear), person of the subject of the *wish*-embedding (same coding convention as main subject) and subjunctivity of singular 1st and 3rd person inflections of *to be*: (*was* = not subjunctive; *were* = subjunctive). We also coded for morphological tense-marking errors, i.e., tense inflections that diverge from the grammatical form used by adults in this structural context. Errors were separated into those that lack past-tense marking in the *wish*-complement, i.e., 'present-for-past' (e.g., "I wish I **have** a banjo") or 'other' tense errors (e.g., "I wish we have **gotted** some mail" or "I wish I **be** a sheep"). For all present-for-past errors, we coded whether they were compatible with a 'bare verb usage' which could signal children having dropped *would/could* (e.g., "I wish I <could> do that"). If a child used an auxiliary ("I wish we **can** eat") or other inflected form ("I wish I'**m** already at home") we marked the error as incompatible with bare verb usage. As a first semantic variable, we coded for the temporal orientation of the embedded clause (e.g., "I wish I had a train" = present; "I wish I had gone to the train" = past; "I wish I would have a train" = future; "I wish want a train" = unclear). Unlike adults, who use *would* in future wishes (e.g., "I wish you **would**n't do that"), children's utterances sometimes lack *would* in wishes with a future temporal orientation (e.g., "I wish you stop bug me"). Since lexical aspect contributes to the temporal orientation (Iatridou, 2000), wishes without *would* were coded as present when containing stative verbs (i.e., *had, was, knew*) and as future when containing eventive verbs (e.g., *go, stop, got*). The tests used to determine stative or eventive lexical aspect came from (Dowty, 1986).

When children use *wish*-constructions, it is not assured that they understand that the *wish* statement is a counterfactual utterance, and thus indicates desires outside one's reach. For this reason, we coded for the evidence we have available as coders to determine whether the wish is used counterfactually or not. We inspected the discourse and situational context as available in CHILDES transcripts, to determine whether the wish demonstrated 'clear' counterfactual reasoning. Counterfactual wishes were considered to contain clear counterfactual reasoning when lexical material within the utterance itself contrasted the actual world with a counterfactual one (e.g.,: "I wish I asked for toast **instead**" = lexical contrast, "I wish you did**n't** do that" = contrast induced by negation, "I wish I **had gone** to the station" = contrast induced by undoing

past event), when the wish desired some sort of existential change, i.e., was counteridentical (e.g.,: "I wish I was a monkey"), or when the utterance was in clear contrast with prior context (e.g.,: "I wish I had green eyes." = contextual contrast when used in a context where it is clear the speaker does not have green eyes). Wishes that were indistinguishable from a regular desire usage (e.g., "I wish I had that horse" or "I wish you'd stop") were marked as having no evidence for counterfactuality, and wishes that were transcribed without context were coded as "inconclusive". Different than for children, we did code adult *wish*-utterances expressing desires such as "I wish I had a kitty" or "I wish I could talk to her" as contextual counterfactuals (without investigating the context it was uttered in) assuming adults always use *wish* counterfactually.

All data was coded by the first author (a fluent non-native speaker). A random subset of 100 child wishes were double-coded, by a native speaker of English (both coders were trained in semantics). An inter-rater reliability analysis was performed to determine consistency among raters in coding for the described variables, using overall accuracy, Gwet's AC1 coefficient (Gwet, 2008) and Cohen's kappa statistic (Cohen, 1968) to describe agreement confidence. While Cohen's kappa statistic is often used as the default method to determine intercoder reliability, it can underestimate reliability in cases where there is high agreement in unbalanced distributions (Gwet, 2008). Since several of our coding variables are unbalanced (e.g., temporal orientation is overwhelmingly present), AC1 is likely a more stable measurement. The exact values for all three different statistics for our coding are displayed in Appendix S2. The AC1 values for all variables exceeded 0.85 (very good agreement) except for the coding indicating the available evidence for counterfactuality (percent agreement = 61%, AC1 = 0.52, $\kappa = 0.49$), which corresponds to moderate agreement (Landis & Koch, 1977). Since coding involves assessments of grammatical and situational contexts, coders discussed all disagreements and came to a consensus for items where either coder missed contextual or grammatical cues in their original rating. The first coder (who coded the entire dataset) was more accurate and conservative than coder two (who only coded the subset). 19 items were judged in favor of coder 1, and 7 items in favor of coder 2. Of the 7 items judged in favor of coder 2, only 1 item was changed from formerly being judged counterfactual to no evidence for counterfactuality. A subset of 13 disagreements remained where coders diverged and contextual cues could be interpreted in different ways. Again, coder 1 tended to code more conservatively, as 11 of these items were categorized as having no or unclear evidence for context-supported counterfactuality, while coder 2 was willing to consider these utterances as true counterfactuals. The intercoder reliability values for evidence of counterfactuality post-discussion corresponded to very good agreement (percent agreement = 87%, AC1 = 0.84, $\kappa = 0.83$). Altogether, this suggests that the coding of our dataset might error on the side of not categorizing potentially counterfactual wishes as

counterfactual, rather than overestimating the instances of wishes displaying counterfactual reasoning.

Data Analysis

For each coded syntactic and semantic variable, we calculated the total count and percentage of occurrences per condition for children and adults separately. We converted the error data into a binary variable coding for the presence or absence of a present-for-past substitution, and modeled the probability of making present-for-past tense errors with a generalized linear mixed-effect model (GLMM, Baayen et al., 2008). We used the glmer-function from the 'lme4' package available on R to perform our analysis (Bates et al., 2015; R Core Team, 2021). We ran two separate models, one over the complete dataset with the fixed effect of age group (child versus adult) to investigate whether children produced more tense errors than adults, and one over the child data with age in months as a fixed effect, to investigate whether children's age predicts their error rate. For both models, we included speaker identity as a random effect to include the variation found among speakers in the model estimates. Inclusion of a random slope or the addition of corpus identity as a random effect did not improve the fit of our models. The model fit (logit link) was estimated by maximum likelihood using the default setting of LaPlace approximation. To test the contribution of our fixed effects we performed a likelihood ratio test comparing our model and a nested model leaving out the variable of interest. We used the 'DHARMa' package (Hartig, 2022) to test the dispersion of our models, and found no indication of overdispersion, which means that the residual variance of our data was not larger than our fitted models assume.

Part 2: Individual Development of Counterfactual Utterances

Selection Criteria & Pre-processing

To gain more insight into the individual longitudinal development of children, we selected children that produced more than 15 wishes. From the complete dataset, six children fit this criterion: Abe - Kuczaj corpus (Kuczaj, 1977), Adam - Brown corpus (Brown, 1973), Laura - Braunwald corpus (Braunwald, 1971), Mark & Ross - MacWhinney corpus (MacWhinney, 1991) and Thomas - Thomas corpus (Lieven et al., 2009). For these 6 children, we searched for counterfactual conditionals by extracting utterances containing *if* in combination with *would, should* and *could*. We proceeded to compare the emergence and development of their first spontaneous counterfactual conditionals against the development of their *wish*-utterances.

Exclusions

The 6 children with longitudinal data were responsible for 175 of the wishes. For those 6 children, we also extracted 341 conditionals with *would, should* or *could*. We excluded 63 utterances where *if* was used like *whether* and not as an *if-then*-conditional (e.g., "see if you could throw two dinosaurs in"), and 93 utterances that did not contain past tense inflection in the *if*-clause. We did this to exclude (non-counterfactual) hypothetical conditionals such as "Maybe you shouldn't be there, if you scare Ellen" or "What would the toilet be like if you flush it?". A total of 185 conditionals remained. Because we were interested in the relative onset difference between counterfactual wishes and conditionals, we decided to be conservative in our inclusion criteria of what consists as a counterfactual. For this reason, we excluded all wishes and conditionals that have future temporal orientation, since their status as counterfactual is debated (strictly speaking, the future cannot be counter-to-fact, as it has not yet occurred). We excluded 26 wishes like "I wish that you stop talking" and 80 conditionals like "Mom what would happen if I taked this balloon". We were left with 104 counterfactual conditionals and 149 wishes with present or past temporal orientation.

Coding Conventions

For the conditionals, we coded for the same semantic variables as we did for the wishes. For temporal orientation this included the categories 'present' (e.g., "they could fly if they had wings") and 'past' (e.g., "what would have happened if they didn't invent houses"). For evidence for counterfactuality this again included clear lexical counterfactuals (e.g., lexical contrast: "only if Super Man was **real** he could do it", negated contrast: "but if I was**n't** a chair I wouldn't be a chair", or past contrast: "yeah it could have lived if I would **have gotten** enough food for all of them"), counteridenticals (e.g., "if I were you I would eat food") or contextual counterfactuals (e.g.,: "if there were four one would hafta wait his turn", when used in a context where there are less than four). Conditionals that were indistinguishable from a regular hypothetical by contextual cues (e.g., "if I could get my boots on I could go inside") or uttered out of context were marked as "inconclusive". Since we excluded all conditional utterances that had present tense marking in the *if*-clause, we could not code for possible present-for-past substitutions.

Control Comparison

We hypothesized that present-for-past substitutions in the *wish*-complement could indicate children have not yet figured out that counterfactual utterances require the "fake" past morphology. Alternatively, it could be the case that some children have

yet to develop the ability to use the past tense in appropriate contexts, and generally avoid using the past tense in any environment, including (but not limited to) counterfactual utterances. To investigate this possibility, we determined for each child the period in which they made present-for-past tense errors and extracted all utterances containing the word *yesterday* during this period, as well as all utterances containing a past tense morpheme. This yielded 29 utterances with *yesterday*, and 7033 utterances with past tense. We looked for signs of productive tense marking by indicating whether children correctly inflected the main verb of utterances containing the temporal adverb *yesterday* with past, and whether their other past utterances included any instances of overregularization (e.g., "I telled daddy something").

Data Analysis

For the coded semantic variables, we calculated the total count and percentage of occurrences per condition for all six children. We created a new variable for evidence of counterfactuality that grouped evidence into binary bins as either "clear" (lexical, counteridentical or contextual evidence) or "unclear" (inconclusive or no evidence). We then compared per child the onset of wishes and conditionals per category, and calculated the difference between the two. We then averaged over children to get an idea of the average difference between the onset of wishes and conditionals. Since we only had data for six children, we discuss these results descriptively and conducted no further statistical analysis.

Results

Part 1: Children's and Adult's Wishes and the "Fake" Past Tense

In total we found 349 *wish*-constructions (*wish* + proposition) in children between the ages of 2 and 6. The first instance of the *wish*-construction we found at 25 months (12a). Like most early wishes, this wish expresses a desire about something mentioned or in direct proximity, e.g., wishing for a horse when looking at horses (12b).

Early Wishes (Like Desires)

(12a) Laura (2;1): I wish I had sandals. (Braunwald, 1971)
(12b) Becky (2;7): I wish I had a horsie. (Manchester: Theakston et al., 2001)

From these early uses, it is not clear whether children know that *wish* can only be used counterfactually, i.e., the desire is unlikely to be fulfilled. So it could be that children initially use *wish* like the regular desire verb *want*. Consistent with this possibility, we

sometimes encounter clear non-counterfactual wishes, where parents comment on the incongruency (13a/b).

Non-Counterfactual Wishes

(13a) Emily (2;1): but I wish that my cold is better. (Nelson, 1989)
Father: yeah you had no cold at all everything's fine.

(13b) Laura (3;2): I wish you were my mommy. (Braunwald, 1971)
Mother: I am your mommy.

For this reason, we coded for the evidence we have available as researchers to believe that a child's wish is produced with a counterfactual meaning in mind. We separated the wishes into 5 categories: wishes that seem clearly counterfactual based on lexical information inside the utterance (14-16), i.e., contrasting the actual world against the postulated one through undoing the past, negation or a lexical contrast (n=43, 12% of total wishes); wishes that indicate an existential change (17), i.e., counteridenticals (n=27, 7.8%); wishes that are in clear contrast with reality as deduced from the discourse context (18) (n=96, 27.5%); wishes that provide no evidence for counterfactuality (n=69, 19.8%) and wishes that are not interpretable without more context and therefore provide inconclusive evidence (n=114, 32.7%).

Clear Evidence for Counterfactuality

Lexical Evidence: Undoing Past

(14) [hearing train in distance] (Thomas: Lieven et al., 2009)
Thomas (3;1): I wish **gone** Burnage Station watch that train.
<later in recording Thomas comments "I'm missing all the trains">

Lexical Evidence: Negation

(15) [mother about to braid child's hair] (Hall: Hall & Tirre, 1979)
Mia (4;9): I wish you did**n't** hafta braid it.

Lexical Evidence: Lexical Contrast

(16) [child pretends it's his birthday] (Thomas: Lieven et al., 2009)
Thomas (4;2): Oh I wish it was my birthday today **really**.

Counteridentical (Change of Identity)

(17) Ross (4;2) I wish humans were **not** humans. (MacWhinney, 1991)

Contextual Evidence

(18) Father: You don't see bumblebees in the dark at all.
Mark (5;10) I wish that the lights were on. (MacWhinney, 1991)

Most wishes uttered by two-year-olds lack clear evidence for counterfactuality. The first *wish*-constructions that we coded as having clear evidence for a counterfactual intended meaning start around 35 months, this is true for all three categories (lexical, counteridentical and contextual). This finding is visually displayed in Figure 1 below.

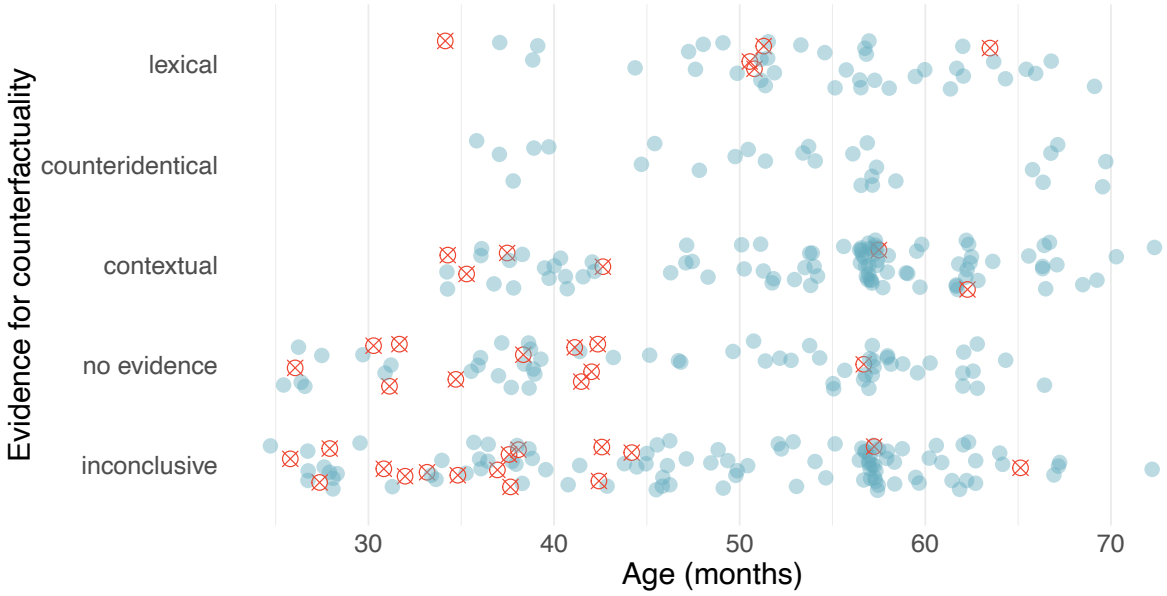


Figure 1. ***Breakdown of children's wishes. Plotted are all children's*** **wish-*productions (N=349) per evidence category for indicating counterfactuality (y-axis). Evidence that is lexical, counteridentical or contextual is considered to indicate clear counterfactuality, while no or inconclusive evidence indicates that it's unclear whether the utterance is used counterfactually. Red struck-through instances indicate the wish contained a present-for-past substitution (e.g., "I wish I have a horse"). The x-axis indicates the speaker's age in months.***

Do Children Produce Wish-Constructions Lacking the "Fake" Past Tense?

To investigate our first question about children's acquisition of the "fake" past tense, we analyzed the tense children used in the complement of the *wish*-constructions. The tense expression in the complement of children's produced wishes diverged from the adult-form in several ways. The most frequently occurring error (38 instances,

10.9% of total), was that of using present tense in the *wish*-complement rather than past tense. For adults, we only documented 4 instances where present tense was used inside the *wish*-complement (0.8% of the total amount of 465 adult wishes). Children are thus not matching their input when making these productive tense substitutions. We modeled the presence or absence of present-for-past errors with a generalized linear mixed-effects model (GLMM) including speaker identity as a random factor to investigate whether age group (child or adult) was a predictor of error rate. A likelihood ratio test comparing our model against a nested model without fixed effects, found that age group was a significant predictor of error rate ($\chi^2(2) = 4.75$, $p = .029$). The odds of making a present-for-past substitution increased for children compared to adults ($\beta = 17.5$, $z = 3.67$, CI = 3.79 - 80.7). Children's present-for-past errors are marked on Figure 1 with red crossed circles. For 15 of these errors, it is not entirely clear whether they are marking present tense or are the consequence of dropping 'would', since the present tense is indistinguishable from bare verb usage in these cases (19). For the remaining 26 errors it was clear that they indicated present tense, i.e., due to inflection (20a) or the choice of auxiliary (20b).

Present-for-Past Errors

(19) Adam (5;2): I wish I **have** a banjo like dis [this]. (Brown, 1973)

(20a) Sarah (3;6): I wish it'**s** valentine. (Brown, 1973)
(20b) Martin (3;6): I wish I **can** be on the tellie. (Wells, 1981)

Present-for-past errors are more common among younger children, especially those between age 2 and 3. With a second GLMM analysis considering speaker identity as a random effect, we confirmed that age in months is a predictor for children's error rate ($\chi^2(2) = 22.26$, $p < .001$). The odds of making a present-for-past mistake decreased with every month ($\beta = .911$, $z = -4.27$, CI = .088 - .951). When we group the present-for-past tense mistake counts by age group (per year) we observe indeed that most present-for-past substitutions occur before age four, and then drop off steeply. This decrease in error rate is displayed in Table 1.

Table 1. ***Count and percentage of present-for-past tense per age window***

Age Group	# children	# wishes	# errors	% of total
2-3	18	47	15	31.9
3-4	21	84	14	16.7
4-5	41	148	6	4.05
5-6	19	70	3	4.29
Total	99	349	38	10.9

Is Usage of the "Fake" Past a Prerequisite for Expressing Counterfactuality?

As can be observed in Figure 1, present-for-past errors were found in wishes for which we have no or inconclusive evidence that the wishes are used counterfactually (11 errors), as well as in wishes that were used in a context that was clearly counterfactual (27 errors). This suggests that the counterfactual's "fake" past is not a necessary component of the *wish*-construction.

Other Tense Errors

Besides making present-for-past errors, we also found that children sometimes express wishes about the past without using the past perfect (21a/b). A similar omission of the *had* auxiliary in the past perfect could be observed in example (14). Interestingly, we observed the same for adults (22).

(21a) Abe (4;4): Are we having pork chops for dinner? (Kuczaj, 1977)
Mother: Yes, that's what you asked for.
Abe (4;4): I wish I **asked** for toast instead.

(21b) [child did not have a nice time at his grandma's] (Thomas: Lieven et al., 2009)
Thomas (3;2): because I wish Mum **come** there.
Investigator: ah, did you miss your mum?

(22a) Mother: oh don't we wish we **had** that three weeks ago
(22b) Mother: don't you wish you **had** them when you were little
(Dickinson & Tabors, 2001)

Comparing Children and Adult's *Wish*-utterances

To gain more insight into the overall properties of children's *wish*-productions compared to their input, we compared the syntactic and semantic properties of the 485 adult and 349 child wishes. The proportion of child wishes (0.02% of all utterances) was overall comparable to the proportion of adult wishes across all corpora (0.03%), and we found that children and adults used wishes in a comparable way (Figure 2). The lion's share of wishes are produced from a 1st person perspective, and children use 1st person main clause subjects (83.7%) even more than adults (76.8%) (Figure 2A). This is compatible with the intuition that young children mostly talk about themselves. Similarly, their wishes are mostly about themselves as well, i.e., the embedded subject is first person (49.3%). In contrast, the embedded subject of adult wishes is balanced for person: 1st (36.3%), 2nd (31.0%) or 3rd (32.3%) person (Figure 2B). As for temporal orientation, we see that both children and adults mostly wish about the

present (children: 76.2%, adults 62.6%), followed by the future (children: 11.7%, adults: 24.9%) or the past (children: 4.0%, adults: 12.3%) (Figure 2C). However, it is possible that the counts for children's past and future wishes are somewhat underestimated, as they sometimes left out the past perfect *had* and future *would* auxiliary (discussed in prior section), making them hard to distinguish from the present (e.g., "I wish I come"). Below you find examples of wishes with present (23), past (24) and future (25) temporal orientation produced by children and adults. Counterfactual wishes with a future orientation often indicated a desire to change a habit or a future event that that has already been planned or whose outcome is determined (23a). The counterfactuality in these cases is the implication that this desire is unattainable. For adults, most of the future-oriented wishes express indirect requests (23b).

Wishes with Present Temporal Orientation

(23a) Ross (5;7): I wish you were a little kid then you would understand. (MW, 1991)
(23b) Mother: I wish it was real money. (Thomas: Lieven et al., 2009)

Wishes with Past Temporal Orientation

(24a) Abe (4;3): I wish we haven't come here. (Kuczaj, 1977)
(24b) Father: Boy, I wish Dallas had won the football game. (Kuczaj, 1977)

Wishes with Future Temporal Orientation

(25a) Matthew(4;7): I wish they'd give ya a fork instead of a spoon. (Gathercole, 1980)
(25b) Father: I wish you'd stop hitting. (MacWhinney, 1991)

When we break down the type of available evidence for counterfactuality, we see that children and adults also pattern alike. Most wishes were judged to be clearly counterfactual based on contextual evidence (children: 27.5%, adults: 47.1%), followed by lexical evidence (children: 12.3%, adults: 19.8%) and counteridenticality (children: 7.7%, adults: 2.6%) (Figure 2D). The fact that we observe less contextual wishes for children than for adults could be a consequence of the fact that we conservatively coded for desire-like wishes in children (e.g., "I wish I had a horse" without clear supporting contextual evidence for counterfactuality was coded as having "no evidence") while we assumed that adults use these wishes as true counterfactuals. Last, we compared the counts of subjunctive usages, by looking at 1st or 3rd person singular conjugations of *to be* in both children (n=54) and adults (n= 67) and coded for whether these were marked with subjunctive (*were*) or not (*was*). We found that adults somewhat rarely used the subjunctive form (19.4%), and for children we observed only 3 instances (5.6%) (Figure 2E). For children, all subjunctive wishes came from the North American corpora. For adults, we found only 2 subjunctive wishes (2.9%) in the United Kingdom corpora. This difference could be due to the fact that our sample

from the North American collection was bigger and skews historically older than our UK-sample. Examples of wishes with and without subjunctive mood are provided below for children (26a/b) and adults (27a/b).

Child Wishes with and without Subjunctive

(26a) David (4;9): I wish I **were** in a car. (Hall: Hall & Tirre, 1979)

(26b) Joey (4;9): Yes, I wish I **was** a spoon. (Hall: Hall & Tirre, 1979)

Adult Wishes with and without Subjunctive

(27a) Father: I wish it **were** but it's not. (Clark, 1979)

(27b) Adult: I'll tell you I wish it **was**. (Hall: Hall & Tirre, 1979)

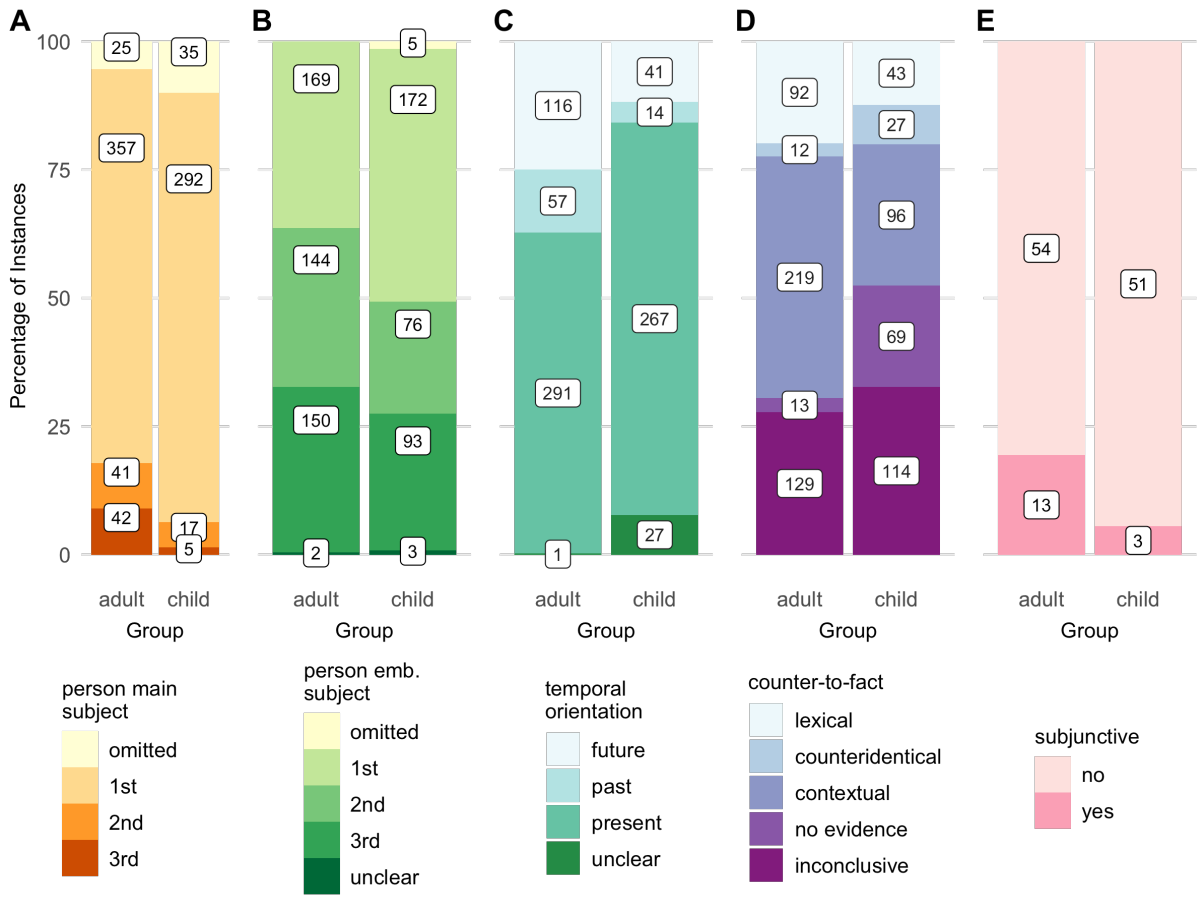


Figure 2. ***Overview of syntactic and semantic properties of child and adult* wish-*constructions. Count (total A-D = 465 for adults and 349 for children, E = 67 for adults and 63 for children) and Percentage (y-axis) of instances.***

Part 2: Individual Development of Counterfactual Utterances

To understand the developmental trajectory of individual children, we investigated the emergence of counterfactual wishes and conditionals in the output of the six children we had enough longitudinal observations for. We investigated both the clarity of the counterfactual (whether there is evidence that indicates the expression is used counterfactually) and whether the child made any present-for-past tense mistakes. The individual development of each child is displayed in Figure 3.

Are Counterfactual Wishes Produced before Counterfactual Conditionals?

The age at which the 6 children started to use the *wish*-construction varied from 2;01 (25 months) to 4;00 (48 months). The age of the first clear counterfactual wish usages fell within a later range between 2;10 (34 months) and 4;11 (59 months). For (both clear and unclear) counterfactual conditionals the onset range was 2;8 (32 months) – 4;4 (52 months). Examples of children's first counterfactual conditional constructions are provided in (28a/b). The onset of the first wish/conditional was often followed with subsequent usages of the constructions within as short period of time. Repeated uses of a new construction within a short period of time is considered to be a signal of productivity (Snyder, 2007; Stromswold, 1990). The first counterfactual wish with past temporal orientation was produced by Thomas at age 3 (29a) and the first counterfactual conditional with past temporal orientation by Abe at age 3;8 (29b). Half the children produced their first past counterfactual construction before the age of 4. All past counterfactual usages are indicated on Appendix Figure S3.

First Counterfactual Conditionals

(28a) Laura (2;8): If a really hole was in here, then I would cry for new pants. (Braunwald, 1971)

(28b) Mark (3;7): We could fly if we had wings well, we don't so we can't, but I know one way how you can fly (MacWhinney, 1991)

First Past Counterfactuals

(29a) Thomas (3): Your wish you gotten on this train. (Thomas: Lieven et al., 2009)

(29b) Abe (3;8): no he would have smelled really bad if he died (Kuczaj, 1977)

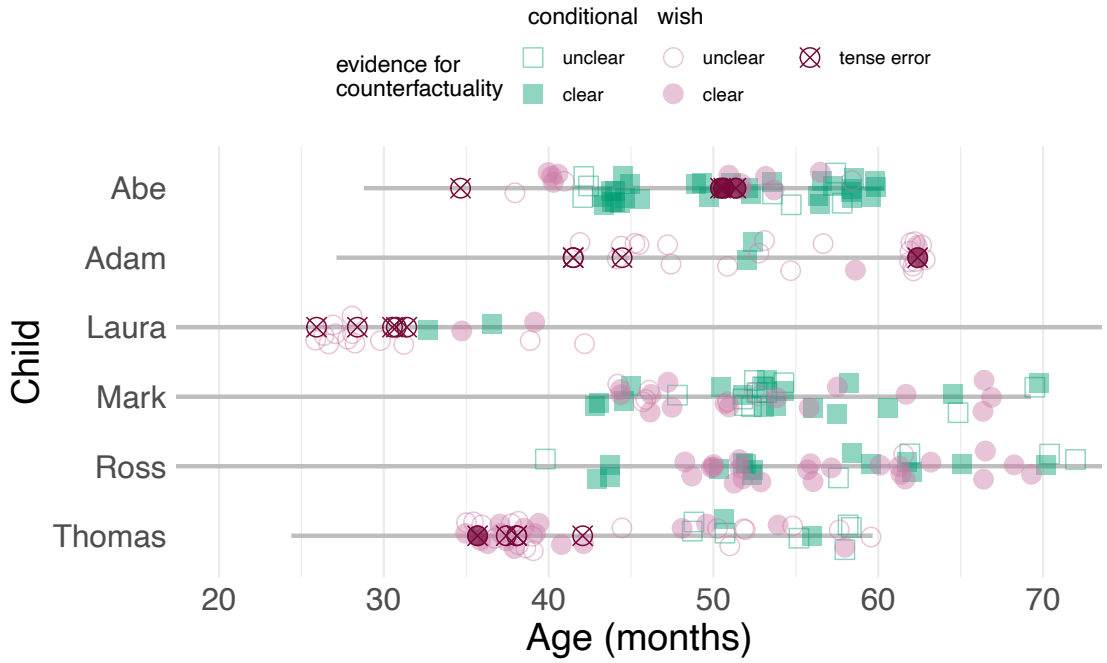


Figure 3. ***Counterfactual conditionals (green squares) and wishes (pink circles) for each child (y-axis) with age indicated in months on the x-axis. Filled shapes indicate that the evidence for counterfactuality is clear, empty shapes indicates the evidence is unclear. Struck-through wishes indicate they contained a tense error in the form of a present-for-past substitution. Grey line indicates recording span. See Appendix S3 shows which of these wishes were used with past temporal orientation.***

To quantify the average difference between the onset of wishes and conditionals for each child, we compared the onset per evidence category (unclear and clear) and calculated the average values. This numerical comparison is displayed in Table 2. On average, children started producing counterfactual wishes before conditionals, though the difference is more prominent if we consider unclear counterfactuals (4.7 months earlier) than if we compare the average onset of clearly counterfactual constructions (0.6 months earlier). However, there is a lot of individual variation in the presence and size of the gap between the onset of the two constructions. 4/6 children start using (unclear) counterfactual *wish*-constructions before they use conditional constructions (difference ranging from 6.6 – 13.6 months), Mark started using both constructions around the same time, and Ross was the only child who used counterfactual conditional constructions before wishes. Comparing clear counterfactual wishes and conditionals, we find that only 2 children (Abe and Thomas) start using wishes before conditionals (difference 3.6 and 15.6 months). For Mark and Laura they

emerge around the same time, and for the last 2 children it seems that clear counterfactual conditionals precede the onset of clear counterfactual wishes (for Adam by 6.4 months, and for Ross by 5.2 months).

Table 2. ***Overview of children's age (in months) at time of first (clear) counterfactual wishes and conditionals (cond.)***

Child	Age 1st wish	Age 1st cond.	Age 1st cond - wish	Age 1st clear wish	Age 1st clear cond.	Age 1st clear cond. - wish
Abe	34.7	42.1	7.4	39.9	43.5	3.6
Adam	41.5	52.4	10.9	58.8	52.4	-6.4
Laura	25.8	32.4	6.6	34.6	32.4	-2.2
Mark	44.6	42.8	-1.8	44.6	42.8	-1.8
Ross	48.3	39.9	-8.4	48.3	43.1	-5.2
Thomas	35.5	49.1	13.6	35.5	51.1	15.6
Average	38.4	43.1	4.7	43.6	44.2	0.6

Present-for-Past Errors

We observed that most present-for-past tense errors occur in the early stages of the emerging *wish*-construction, regardless of the age the child started using the construction. It should be noted again that we found present-for-past errors in both unclear (n= 13) and clear (n= 5) counterfactual wishes. Two children (the siblings Mark and Ross) never made a present-for-past substitution in their wishes, and two children (Laura and Thomas) made multiple present-for-past substitutions when they started using the *wish*-construction, and then stopped making them before their first counterfactual conditionals emerged. This means that for 4/6 children, present-for-past substitutions did not occur after the onset of the counterfactual conditional. Adam and Abe complicate this picture. Adam initially stopped making tense errors around 45 months (about 7 months before his first counterfactual conditional), but then slipped up at age 5;2 (62 months). Since this also marked the end of his recording period, it is unclear whether he made any more present-for-past substitutions after this occurrence. Abe is unique in making present-for-past substitutions when both his counterfactual wishes and conditionals are productive (at age 4;3, 51 months).

Productive Tense Marking

Lastly, we examined children's overall productive past tense usage during the period where they made present-for-past errors in counterfactual constructions. We did this to investigate whether their present usage in counterfactual contexts is due to a

variable or inconsistent use of past tense marking in general. For each child, we recorded the successful and unsuccessful instances of past tense marking in the context of the temporal adverb *yesterday*, and the period over which they exhibit overregularization. This is displayed in Figure 4. For all children, we found indications of productive past tense usage (both from overregularization and past tense usage with *yesterday*) outside counterfactual contexts during their error period. While Abe used present inflection once in a *yesterday* utterance at the onset of his error period, he later correctly started using past tense in this environment. For Laura we found multiple present tense errors with *yesterday* before 28 months. This indicates that some of Abe's and Laura's earliest errors could be due to a general immature use of the past tense.

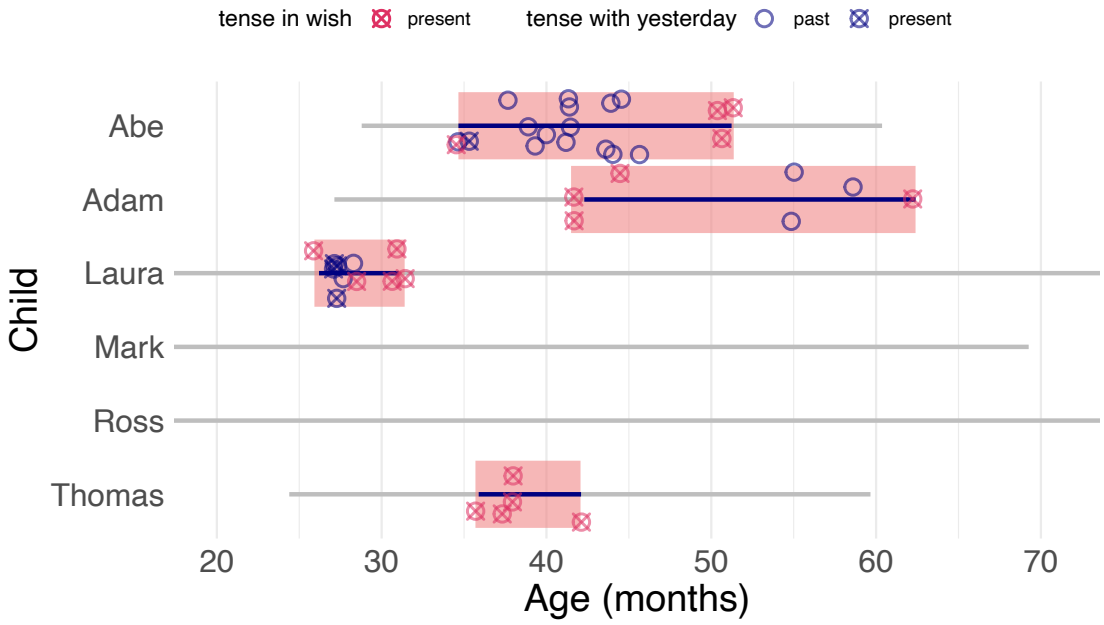


Figure 4. ***Overview of children's productivity with the past tense. Pink rectangles indicate the time span in which individual children (y-axis) produced wishes with tense errors. Each instance of a present-for-past error in wishes is displayed as a pink crossed circle. Within the error span, we plotted the tense of utterances with yesterday with blue circles (crossed means present tense was used). Blue lines within the error span indicate the time span over which we found instances of overregularization (e.g., "I putted"). Grey line indicates recording span. See Appendix S4 for corresponding numeric information in table format.***

Discussion

In this paper we examined the first language acquisition of counterfactual utterances, with our main focus on the development of children's wishes. We conducted corpus research that consisted of two parts. First, we extracted all child and adult utterances containing the word *wish* from eligible corpora on CHILDES and coded for various syntactic and semantic variables. We provided a detailed overview of children's *wish*-constructions and compared the properties of *wish*-utterances produced by children and adults. Second, we took a closer look at the longitudinal linguistic development of 6 children and investigated the maturation of their counterfactual language, comparing their usage of counterfactual wishes and conditionals. With this research we addressed two questions related to form-to-meaning mapping. First, we asked whether children go through a stage where they map the counterfactual's "fake" past morpheme to actual past temporal orientation, and consequently generate present tense inflected verbs in their own productions of present counterfactual constructions. Second, we asked whether linguistically more transparent counterfactual constructions (wishes) are acquired before the more complex counterfactual conditional. The combined results of our corpus work show there are indeed children that go through a stage where they productively use present tense in the complement of counterfactual wishes, diverging from their adult input. We also found that the average age children start using wishes is 3;2 (onset ranging between 2;1 and 4;0), which is before the average onset of counterfactual conditionals around age 3;7 (range between age 2;8 and 4;4). These general findings are compatible with the view that linguistic transparency plays a role in the acquisition of counterfactuality. However, the longitudinal data also illustrates that each child has a unique developmental trajectory, which leads to differences in when individual children start speaking counterfactually and which constructions they initially use. Below we discuss our questions and findings in more detail, as well as limitations to this work and suggestions for future research.

Children's Counterfactuals Contain Present-for-past Errors

The first question addressed in this study was whether children go through a phase where they make tense-marking mistakes in the complement of counterfactuals. Acquiring counterfactual utterances requires discovering that the past tense in its complement/antecedent is "fake" and marks counterfactuality instead. This mapping between counterfactuality and the past tense morpheme is thought to require complex semantic operations (Iatridou, 2000; Karawani, 2014; Ritter & Wiltschko, 2014). Since children have to see through the "fakeness" of the past tense in order to learn this mapping, we hypothesized that children would productively form counterfactual

wishes that have a present tense (rather than past tense) marking on the embedded matrix verb, as this aligns with the temporal orientation of a present wish. Indeed, we found that children make a substantial amount of past tense errors (11% of total wishes), most of them between ages 2 and 4 (75.6%). We observed these errors both in wishes that were judged to have clear evidence for a true counterfactual usage, and in wishes that were less clearly adult-like for counterfactuality. The fact that we observed present tense in clear counterfactual wishes, suggests children do not need the "fake past" to express counterfactual meaning. Instead, it's possible they mapped counterfactual meaning directly to the verb *wish*. The fact that you can express counterfactual meaning without relying on the "fake" past is consistent with cross-linguistic typology for counterfactual constructions: there are languages that express counterfactuality without making use of tense-marking, e.g., Mandarin Chinese (Jiang, 2019; Yong, 2016). This is also consistent with the fact that we observed some past counterfactuals productions with only one layer of past marking (21/22).

One could wonder whether the tense errors found in the complement of *wish* could be due to children not yet having acquired the past tense form in general. This seems unlikely, as children generally have productive past tense usage before age 3 (Brown, 1973; de Villiers, 2000; Kuczaj, 1977). For example, Abe acquired past tense with a 90% success rate by age 2;9, right before his first counterfactual wishes occurred (Kuczaj, 1977). For three children, we showed that they display clear signs of productive tense marking during the period in which they make tense marking errors in counterfactual constructions. They use past tense in utterances with *yesterday* and overregularize the past tense morpheme to irregular verbs, showing productive usage. Only for the youngest *wish*-producer, Laura, do we find some tense marking errors outside counterfactual constructions, suggesting that her earliest errors (before 28 months) might be partially due to a general problem with applying past tense inflection. Another explanation for present-for-past tense errors could be that children actually use a bare verb construction (rather than present tense) because they treat *wish* analogously to the semantically related desire verb *want* (which selects for a non-finite complement). Or they may be omitting the auxiliary verb *would* in future wishes, which is plausible as it is often pronounced in reduced form. However, from the 41 errors only 15 (37%) are compatible with a bare verb/dropped *would* explanation, which suggests that this cannot be the sole reason for children's past tense errors. Most tense errors in wishes are thus due to productive present tense marking, counter to the examples children receive in their input.

Children's Start Producing Wishes before Conditionals

The second aim of this corpus study was to find out whether counterfactual wishes are acquired before counterfactual conditionals. Since *wish* is a dedicated marker of counterfactuality in English when it associates with propositional content, we hypothesized that counterfactual wishes would be easier to acquire than counterfactual conditional constructions. Indeed, we found that children generally produced the *wish*-construction either before or simultaneously with counterfactual conditionals. Counterfactual wishes mostly seem to emerge between age 2 and 4, while counterfactual conditionals emerge between age 2.5 and 4.5. However, it should be noted that there is a wide range of variation between children and the presence and size of the gap between the onset of wishes and conditionals. Some children acquire wishes before conditionals with an onset gap ranging from half a year to a year, while other children start using both constructions around the same time. We also indicated the need to be cautious not to equate using the *wish*-construction with having the ability to reason counterfactually about the world. Indeed, children's early wishes do not always seem adultlike. Especially children under age 3 seem to use the *wish*-construction to express direct desires (much like the verb *want*), and it is unclear whether they know *wish* can only be used when you believe this desire to be counterfactual. We start finding clear indication of wishes with unequivocal counterfactuality (based on contextual and lexical information) between age 2.5 and 5, and for counterfactual conditionals this range is 2.5 to 4.5. While some children's samples display a long gap between using clear counterfactual wishes and conditionals (ranging from 3-16 months), other children's samples use clear counterfactual conditionals before wishes (difference ranging from 2 to 6 months). However, it should be noted that the distinction of "clear" versus "unclear" completely relied on the coder's interpretation. As discussed before, the coding was done conservatively to reduce the chance of overinterpreting the counterfactuality of an utterance, which thus means we might be underestimating the counterfactuality of utterances we deemed "unclear". If we take our findings at face value, however, they suggest that the *wish*-construction is generally acquired before or simultaneous with the counterfactual conditional. While it's not clear whether children always use the construction in an adultlike way, at least some children also display this pattern in the onset of clear counterfactual wishes and conditionals.

Crucially, it is unlikely that the difference we observe between the acquisition of counterfactual wishes and conditionals is solely due to the difference in causal structure (i.e., the *if…then* relationship in conditionals). While intuitively, conditionals are harder to process because they rely on linking two clauses with a causal relation, we actually find that most children start producing the non-counterfactual conditional structure (e.g., hypothetical future) before age 3 (Kuczaj & Daly, 1979; Reilly, 1982).

Since most children start producing wishes after age 3, the difficulty of the conditional structure itself is not holding them back from acquiring the counterfactual conditional at that time. Another question that might arise is how accurate the ages of acquisition are that we found for the different constructions. Since corpus data is sampled and only includes a small proportion of the actual spoken input and output of the child, there is always the risk that we have missed earlier occurrences of either the wishes or conditionals. However, since the density of the used corpora was high (recording 1-5 times a month), the sample size of the observed constructions fairly similar (we observed 149 wishes and 104 conditionals) and the onset difference we observed quite large (6 to 12 months), we believe it to be unlikely that the onset differences we observed are solely due to unequal sampling.

Individual Variation

The development of counterfactual language depends on an interplay of different factors, including the development of specific grammatical structures (e.g., the past tense, conditional constructions and embedding), the development of counterfactual reasoning (e.g., thinking about possibilities and keeping in mind conflicting information), the transparency of different constructions and the consistency of children's input. Each of these factors can influence the onset of counterfactual constructions in children's speech, and individual variation between children is expected given these different forces that are at play. In this paper, we specifically focused on the role of linguistic transparency on the acquisition of counterfactuality, predicting that the complexity involved with acquiring the counterfactual's "fake" past tense may lead to present-for-past errors in children's early counterfactual productions, and that counterfactual wishes are easier to acquire than counterfactual conditionals. We found evidence supporting these ideas: from the six children we have longitudinal data for, four were found to make productive present-for-past errors and produce wishes before counterfactual conditionals. However, it is important to reflect on the fact that not all children did. In particular, the counterfactual development of the brothers Ross and Mark (MacWhinney, 1991) followed a strikingly similar trajectory to each other that was distinct from the developmental pattern we observed in the other children. Despite their age difference, both children started producing their first counterfactual constructions around age 3.5, both children almost immediately produced these counterfactual constructions in clear adult-like counterfactual situations, both children used counterfactual conditionals before or simultaneously with counterfactual wishes, and both children have not been found to make any present-for-past errors. Perhaps, this similarity can be attributed to their shared genetic make-up and/or the fact that they grew up under similar circumstances, e.g., receiving a comparable amount and quality of speech input. But how come the brothers'

counterfactual development differs from that of the other children in our sample? One possibility is, that Ross and Mark were somewhat precautious learners that only started using counterfactual constructions once they figured out the exact meaning and mapping (à la Snyder, 2007). Linguistic transparency may have played a role in their early counterfactual development behind the scenes, but any form-to-meaning mapping difficulties were resolved by the time they actually started using these constructions in their own speech. This could explain why the brothers started using counterfactual constructions fairly late compared to some other children, as well as why they immediately started using their counterfactual constructions with an appropriate use of the "fake" past tense in clear adult-like counterfactual contexts. Alternatively, it could be that the brother's input contained particularly salient examples of counterfactual constructions being used in counterfactual situations, facilitating the form-to-meaning mapping task from the beginning, or that cognitive factors were at play. Possibly, the brothers developed the cognitive ability to reason counterfactually after the linguistic mechanisms underlying counterfactual constructions were already in place, while other children developed counterfactual reasoning abilities before they fully acquired the linguistic structures supporting counterfactual language. In the next section we discuss the interplay between linguistic and cognitive complexity in some more detail.

Untangling Linguistic and Cognitive Complexity

As discussed thoroughly in the introduction, the acquisition of counterfactuality relies on both linguistic and cognitive development. On the one hand, children need to develop a concept of counterfactuality and the cognitive abilities to support counterfactual reasoning. On the other hand, children need to acquire the linguistic structures that express counterfactuality in their language, and map counterfactual meaning onto these linguistic expressions. Can we untangle the influence of cognitive complexity and linguistic complexity in the acquisition of counterfactuality? In this study, we showed that children start producing present counterfactual wishes and conditionals as early as age 2, which corresponds to early observations by Bowerman (1986). However, we also noted that children only start using these constructions in contextually salient counterfactual contexts around age 3, suggesting that these initial constructions might precede the concept of counterfactuality. At age 3, children also start producing counterfactual wishes and conditionals about the past, although their productions are not adult-like, lacking the past perfect construction.

From corpus data alone, we cannot know whether children have acquired the ability to reason counterfactually at this age, but the way they use counterfactual constructions spontaneously are suggestive that they do. Why then, do 3-year-old children

often fail counterfactual comprehension tasks? While comprehension research often reports that 4-year-olds, but not 3-year-olds have developed the ability to reason counterfactually (Guajardo et al., 2009; Nyhout & Ganea, 2019; Riggs et al., 1998; Robinson & Beck, 2000), this type of research mostly considers past counterfactual conditionals. Is it possible, that children struggle with the past construction specifically, rather than counterfactual reasoning itself? Our results suggest they might, three-year-olds spontaneously use counterfactual constructions undoing past events, but not yet using a past perfect, e.g., "No he would have smelled really bad if he **died**". In fact, we found the same pattern in adults, a phenomenon that has been extensively described by Crutchley (2004, 2013). Even adults, sometimes use a single past marker for counterfactuals with past temporal orientation, instead of the double past marking, e.g., "If they **took** my wages into consideration, they would have let us buy next door even" (Crutchley, 2013, 15). In fact, the canonical 'past counterfactual construction' only accounted for one third of the variety of structures adult speakers used to talk counterfactually about the past (Crutchley, 2013, p. 456). This variability, in combination with the fact that past counterfactuals are a lot less common than present counterfactual constructions in spontaneous speech, does suggest the linguistic complexity of the past counterfactual construction could contribute to children's difficulty understanding these types of constructions. However, this idea requires future exploration.

Bootstrapping of the "Fake" Past Tense

When looking at the longitudinal data of six children we observed a noteworthy, yet unreliable pattern we will speculate about. For 4/6 children, present-for-past substitutions did not occur after the onset of the counterfactual conditional. For half of them, this was simply because they were never observed making any present-for-past errors. This finding is compatible with a scenario where children first start to use the counterfactual *wish*-construction without having discovered the relation between the "fake" past and the expression of counterfactual meaning. Then, once children successfully figure out this mapping, they cease using the present tense in wishes. Since they have now acquired the mapping between "fake" past and counterfactuality, they can start observing it in other environments, i.e., the counterfactual conditional, allowing them to attribute counterfactual meaning to the conditional construction as well. In other words, it is possible that the dedicated *wish*-construction in English bootstraps the acquisition of the "fake" past, which in turn facilitates learning the counterfactual conditional. However, there are children (i.e., Abe and Adam) that do not follow this pattern. Abe starts using the counterfactual conditional before the end of his present-for-past error period. Notably, Abe also participated in a longitudinal study investigating the development of hypothetical conditionals (Kuczaj & Daly, 1979), so this could have accelerated his acquisition of the counterfactual conditionals

compared to other children. For Adam, the recordings ended before we could determine whether his unexpected present-for-past error at age 5 was an unremarkable slip-up or a continuation of his error period. A fully analogous argument has been made for dedicated epistemic adverbs like *maybe* as potentially helping children learn the more complex variable-meaning modal verbs like *may* or *must* (i.e. auxiliaries with both epistemic and deontic (or other root modality) meanings). However, since we only had longitudinal data available for a small subset of children, we cannot draw any hard conclusions from this sample about the bootstrapping hypothesis.

Considerations and Future Directions

In this paper, we have investigated the acquisition of counterfactual constructions from a form-to-meaning mapping perspective and argued that the linguistic complexity of the counterfactual constructions contributes to their relatively late acquisition. The thought that complexity of linguistic structures plays a role in the emergence of such structure in children's speech is by no means original (Cournane, 2021; Reilly, 1982). For example, Reilly summarizes the relationship between cognitive and linguistic complexity as follows: "*Language and cognition are independent yet interactive systems where cognition is basically responsible for the sequence of acquisition, but it's the linguistic complexity of a structure that determines when that structure will appear in a child's grammar*." (Reilly, 1982, p.xi). We view the process of acquiring counterfactual constructions in a similar way. In order to communicate counterfactuality, children need to have reached certain developmental milestones, including the abilities of holding multiple possibilities in mind (Leahy & Carey, 2019) and considering a false possibility temporarily true (Beck, McColgan, et al., 2011; Byrne, 2007). However, the onset of a linguistic construction also depends on various factors, including its linguistic complexity. Specifically, we argue that constructions that are dedicated to expressing counterfactuality (propositional wishes in the case of English) should help children to detect these constructions in their input, and in the case of English, help discover the link between counterfactuality and the "fake" past tense.

In the future, this hypothesis can be tested by doing comprehension studies investigating children's understanding of counterfactual wishes and conditionals, and by looking at other dedicated counterfactual constructions in other languages to compare their acquisition onset with that of multi-purpose constructions. If having a dedicated counterfactual construction (such as the *wish*-construction) indeed facilitates the discovery of the mapping of counterfactual meaning to the "fake" past, we expect this pattern to hold for other languages as well. As mentioned before, the amount of data we extracted was relatively small, considering that we looked through all eligible corpora available on CHILDES. Since the natural occurrence of counterfactual

constructions is fairly uncommon, future research directly targeting questions about "fake" past-tense usage might want to consider an elicitation task to elicit counterfactual speech, especially when working with languages that have relatively little (or no) corpus data available.

Conclusion

All in all, our findings are compatible with the view that counterfactual constructions are not only challenging because they require complex reasoning, but also because they involve complex form-to-meaning mapping. First, we showed that the counterfactual's "fake" past tense is a complex component of the English counterfactual construction, and that present-for-past tense errors occur in children's speech suggesting that children's initial representation of counterfactual wishes does not always include the obligatory "fake" past marking. However, these non-adult-like productions appear in appropriate counterfactual contexts, suggesting that the "fake" past is not a necessary prerequisite for expressing counterfactuality. Second, we found evidence that children generally acquire the more transparent counterfactual *wish*-construction before counterfactual conditionals. Studies solely focusing on the acquisition of counterfactual conditionals might thus underestimate children's ability to engage in counterfactual reasoning, confounding cognitive with linguistic complexity. However, these results are based on limited data and require larger consideration of the issue. Future research should investigate what role linguistic complexity plays in children's comprehension of counterfactual constructions, as well as how dedicated and undedicated counterfactual constructions are acquired in other languages.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Beck, S. R. (2016). Why What Is Counterfactual Really Matters: A Response to Weisberg and Gopnik. *Cognitive Science, 40*(1), 253–256. https://doi.org/10.1111/cogs.12235

Beck, S. R., McColgan, K. L. T., Robinson, E. J., & Rowley, M. G. (2011). Imagining what might be: Why children underestimate uncertainty. *Journal of Experimental Child Psychology, 110*(4), 603–610. https://doi.org/10.1016/j.jecp.2011.06.010

Beck, S. R., Riggs, K. J., & Burns, P. (2011). Multiple Developments in Counterfactual Thinking. In *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199590698.001.0001/acprof-9780199590698-chapter-6

Beck, S. R., Riggs, K. J., & Gorniak, S. L. (2009). Relating developments in children's counterfactual thinking and executive functions. *Thinking & Reasoning, 15*(4), 337–354. https://doi.org/10.1080/13546780903135904

Bellinger, D. C., & Gleason, J. B. (1982). Sex differences in parental directives to young children. *Sex Roles, 8*(11), 1123–1139.

Bjorkman, B. M., & Halpert, C. (2017). In an imperfect world: Deriving the typology of counterfactual marking. In *Modality Across Syntactic Categories.* (Vol. 1). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198718208.003.0009

Bliss, L. S. (1988). Modal usage by preschool children. *Journal of Applied Developmental Psychology, 9*(3), 253–261.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology, 6*(3), 380–420.

Bohannon III, J. N., & Marquis, A. L. (1977). Children's control of adult speech. *Child Development*, 1002–1008.

Bowerman, M. (1986). First steps in acquiring conditionals. In E. C. Traugott, A. G. Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), *On conditionals* (pp. 285–308). Cambridge University Press.

Braunwald, S. R. (1971). Mother-child communication: The function of maternal-language input. *Word, 27*(1–3), 28–50.

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

Buchsbaum, D., Bridgers, S., Weisberg Skolnick, D., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1599), 2202–2212. https://doi.org/10.1098/rstb.2012.0122

Byrne, R. M. J. (2007). *The rational imagination: How people create alternatives to reality*. MIT press.

Clark, E. V. (1979). Building a vocabulary: Words for objects, actions and relations. *Language Acquisition*, 149–160.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition: The 20th Annual Carnegie Mellon Symposium on Cognition*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=1619055

Clark, E. V. (2001). Emergent categories in first language acquisition. In M. Bowerman & S. Levinson (Eds.), *Language Acquisition and Conceptual Development* (1st ed., pp. 379–405). Cambridge University Press. https://doi.org/10.1017/CBO9780511620669.015

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213. https://doi.org/10.1037/h0026256

Cournane, A. (2021). Revisiting the epistemic gap: It's not the thought that counts. *Language Acquisition, 28*(3), 215–240. https://doi.org/10.1080/10489223.2020.1860054

Crutchley, A. (2004). 'If She Had of Shutted the Cage, the Rabbit Wouldn't Escape': Past Counterfactuals Elicited from 6-to 11-Year-Old Children. *First Language, 24*(2), 209–240. https://doi.org/10.1177/0142723704044935

Crutchley, A. (2013). Structure of child and adult past counterfactuals, and implications for acquisition of the construction. *Journal of Child Language, 40*(2), 438–468. http://dx.doi.org/10.1017/S0305000912000049

Cruttenden, A. (1978). Assimilation in child language and elsewhere. *Journal of Child Language, 5*(2), 373–378.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research, 38*(6), 1199–1211.

Davis, B., Van der Feest, S., & Hoyoung, Y. (2018). Speech sound characteristics of early words: Influence of phonological factors across vocabulary development. *Journal of Child Language, 45*(3), 673–702.

de Villiers, J. (2000). Language and theory of mind: What are the developmental relationships? In *Understanding other minds: Perspectives from developmental cognitive neuroscience, 2nd ed* (pp. 83–123). Oxford University Press.

Demetras, M. (1989). *Working parents' conversational responses to their two-year-old sons.* [Ph.D.]. the University of Arizona.

Demetras, M. J., Post, K. N., & Snow, C. E. (1986). Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language, 13*(2), 275–292.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech, 49*(2), 137–173.

Dickinson, D. K., & Tabors, P. O. (2001). *Beginning literacy with language: Young children learning at home and school.* Paul H Brookes Publishing.

Dowty, D. R. (1986). The Effects of Aspectual Class on the Temporal Structure of Discourse: Semantics or Pragmatics? *Linguistics and Philosophy, 9*(1), 37–61. JSTOR.

Dudman, V. H. (1983). Tense and time in English verb clusters of the primary pattern. *Australian Journal of Linguistics, 3*(1), 25–44. https://doi.org/10.1080/07268608308599298

Edwards, J., & Beckman, M. E. (2008). Methodological questions in studying consonant acquisition. *Clinical Linguistics & Phonetics, 22*(12), 937–956.

Evans, M. A. (1985). Self-initiated speech repairs: A reflection of communicative monitoring in young children. *Developmental Psychology, 21*(2), 365.

Forrester, M. A. (2002). Appropriating cultural conceptions of childhood: Participation in conversation. *Childhood, 9*(3), 255–276.

Francis, G. A., & Gibson, J. L. (2021). *Pretense, Executive Functions, and Counterfactual Reasoning: Evaluating the Case for a 'Unified Theory of Imaginative Processes.'* PsyArXiv. https://doi.org/10.31234/osf.io/skxb8

Garvey, C., & Hogan, R. (1973). Social speech and social interaction: Egocentrism revisited. *Child Development,* 562–568.

Gathercole, V. (1986). The acquisition of the present perfect: Explaining differences in the speech of Scottish and American children. *Journal of Child Language, 13,* 537–560. https://doi.org/10.1017/S0305000900006875

Gathercole, V. C. M. (1980). *Birdies like birdseed the bester than buns: A study of relational compratives and their acquisition.* [Unpublished PhD dissertation]. University of Kansas.

Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., Pappas, A., & Keil, F. C. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development,* i–157.

Gelman, S. A., Taylor, M. G., Nguyen, S. P., Leaper, C., & Bigler, R. S. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development,* i–142.

Gelman, S. A., Ware, E. A., Kleinberg, F., Manczak, E. M., & Stilwell, S. M. (2014). Individual differences in children's and parents' generic language. *Child Development, 85*(3), 924–940.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard Words. *Language Learning and Development, 1*(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Gopnik, A., & Walker, C. M. (2013). Considering Counterfactuals The Relationship between Causal Learning and Pretend Play. *American Journal of Play, 6*(1), 15–28.

Gopnik, M. (1989). Reflections on challenges raised and questions asked. In P. R. Zelazo, R. G. Barr, & P. D. Zelazo (Eds.), *Challenges to Developmental Paradigms: Implications for Theory, Assessment and Treatment* (pp. 259–273). Taylor

& Francis Group. http://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=1639232

Guajardo, N. R., Parker, J., & Turley-Ames, K. (2009). Associations among false belief understanding, counterfactual reasoning, and executive function. *British Journal of Developmental Psychology, 27*(3), 681–702. https://doi.org/10.1348/026151008X357886

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29–48. https://doi.org/10.1348/000711006X126600

Haggerty, L. C. (1930). What a two-and-one-half-year-old child said in one day. *The Pedagogical Seminary and Journal of Genetic Psychology, 37*(1), 75–101.

Hall, W. S., & Tirre, W. C. (1979). The communicative environment of young children: Social class, ethnic, and situational differences. *Center for the Study of Reading Technical Report; No. 125*.

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-level / Mixed) Regression Models.* (R package version 0.4.5). http://florianhartig.github.io/DHARMa/

Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. Oxford University Press on Demand.

Hicks, D. (1991). Kinds of texts: Narrative genre skills among children from two communities. In A. McCabe & C. Peterson (Eds.), *Developing narrative structure* (pp. 55–87). Hillsdale, N.J. : L. Erlbaum.

Higginson, R. P. (1985). *Fixing: Assimilation in language acquisition* [Unpublished PhD dissertation]. Washington State University.

Iatridou, S. (2000). The grammatical ingredients of counterfactuality. *Linguistic Inquiry, 31*(2), 231–270.

Inkelas, S., & Rose, Y. (2007). Positional neutralization: A case study from child language. *Language*, 707–736.

Ippolito, M. (2006). Semantic Composition and Presupposition Projection in Subjunctive Conditionals. *Linguistics and Philosophy, 29*(6), 631–672. https://doi.org/10.1007/s10988-006-9006-2

Ippolito, M., & Keyser, S. J. (2013). *Subjunctive Conditionals: A Linguistic Analysis*. MIT Press. http://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=3339677

James, D. (1982). Past Tense and the Hypothetical a Cross-Linguistic Study. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language," 6*(3), 375–403.

Jiang, Y. (2019). Ways of expressing counterfactual conditionals in Mandarin Chinese. *Linguistics Vanguard, 5*(s3). https://doi.org/10.1515/lingvan-2019-0009

Jipson, J. L., Gülgöz, S., & Gelman, S. A. (2016). Parent–child conversations regarding the ontological status of a robotic dog. *Cognitive Development, 39*, 21–35.

Johnson, M. G. (1986). *A computer-based approach to the analysis of child language data*. University of Reading.

Judd, A. (2018). *Exploring relationships between phonological awareness and phonological productive abilities of kindergarten-aged children* [Master Thesis, Memorial University of Newfoundland]. https://research.library.mun.ca/13353/

Karawani, H. (2014). *The real, the fake, and the fake fake in counterfactual conditionals, crosslinguistically*. LOT.

Karawani, H., & Zeijlstra, H. (2013). The semantic contribution of the past tense morpheme *kaan* in Palestinian counterfactuals. *Journal of Portuguese Linguistics, 12*(1), 105–119. https://doi.org/10.5334/jpl.79

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 589–600. https://doi.org/10.1016/S0022-5371(77)80021-2

Kuczaj, S. A., & Daly, M. J. (1979). The development of hypothetical reference in the speech of young children*. *Journal of Child Language, 6*(3), 563–579. https://doi.org/10.1017/S0305000900002543

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child* (pp. xi, 250). Harvard University Press.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

Leahy, B. P., & Carey, S. E. (2019). The Acquisition of Modal Concepts. *Trends in Cognitive Sciences, 24*(1), 65–78. https://doi.org/10.1016/j.tics.2019.11.004

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*(3), 481–507.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk. Hillsdale, NJ, US*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Transcription format and programs* (Vol. 1). Psychology Press.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language, 17*(2), 457–472.

Morisset, C. E., Barnard, K. E., Greenberg, M. T., Booth, C. L., & Spieker, S. J. (1990). Environmental influences on early language development: The context of social risk. *Development and Psychopathology, 2*(2), 127–149. https://doi.org/10.1017/S0954579400000663

Nelson, K. (1989). *Narratives from the crib*. Cambridge, Mass. : Harvard University Press.

Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language, 43*(5), 1158–1173.

Ninio, A., Snow, C. E., Pan, B. A., & Rollins, P. R. (1994). Classifying communicative acts in children's interactions. *Journal of Communication Disorders, 27*(2), 157–187.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4- and 5-year-olds. *Cognition, 183*, 57–66. https://doi.org/10.1016/j.cognition.2018.10.027

Ogihara, T. (2000). Counterfactuals, Temporal Adverbs, and Association with Focus. *Semantics and Linguistic Theory, 10*(0), 115–131. https://doi.org/10.3765/salt.v10i0.3106

Parsons, J. M. (2006). *Positional effects in phonological development: A case study*. Memorial University of Newfoundland.

Pater, J. (1997). Minimal violation and phonological development. *Language Acquisition, 6*(3), 201–253.

Peterson, C., & McCabe, A. (1983). *Developmental Psycholinguistics: Three Ways of Looking at a Child's Narrative*. Springer. http://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=3084310

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

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world." *Child Development, 81*(1), 376–389.

Rafetseder, E., & Perner, J. (2012). When the alternative would have been better: Counterfactual reasoning and the emergence of regret. *Cognition & Emotion, 26*(5), 800–819.

Reilly, J., Snitzer. (1982). *The Acquisition of Conditionals in English* [Unpublished PhD dissertation]. University of California.

Rett, J., & Hyams, N. (2014). The Acquisition of Syntactically Encoded Evidentiality. *Language Acquisition, 21*(2), 173–198. https://doi.org/10.1080/10489223.2014.884572

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development, 13*(1), 73–90. https://doi.org/10.1016/S0885-2014(98)90021-1

Ritter, E., & Wiltschko, M. (2014). The composition of INFL. *Natural Language & Linguistic Theory, 32*(4), 1331–1386. https://doi.org/10.1007/s11049-014-9248-6

Robinson, E. J., & Beck, S. (2000). What is difficult about counterfactual reasoning. *Children's Reasoning and the Mind*, 101–119.

Romero, M. (2014). 'Fake Tense' in Counterfactuals: A Temporal Remoteness Approach. *The Art and Craft of Semantics: A Festschrift for Irene Heim, 2*, 47–63.
Rowland, C. F., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language, 33*(4), 859–877.
Sachs, J., & Nelson, K. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language, 4*, 1–28.
Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods, 51*(4), 1928–1941. https://doi.org/10.3758/s13428-018-1176-7
Schneider, P., Hayward, D., & Dubé, R. V. (2006). Storytelling from pictures using the Edmonton narrative norms instrument. *Journal of Speech Language Pathology and Audiology, 30*(4), 224.
Schulz, K. (2014). Fake Tense in conditional sentences: A modal approach. *Natural Language Semantics, 22*(2), 117–144. https://doi.org/10.1007/s11050-013-9102-0
Slobin, D. (1973). Cognitive prerequisites for the development of grammar. *Studies of Child Language Development, 1*, 75–208.
Smith, N. V. (1973). *The acquisition of phonology: A case study*. Cambridge University Press.
Snyder, W. (2007). *Child Language: The Parametric Approach*. OUP Oxford.
Sprott, R. A. (1992). Children's use of discourse markers in disputes: Form-function relations and discourse in child language. *Discourse Processes, 15*(4), 423–439.
Stromswold, K. J. (1990). *Learnability and the acquisition of auxiliaries* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/13715
Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29*(2), 103.
Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*(1), 127–152. https://doi.org/10.1017/S0305000900004608
Tommerdahl, J., & Kilpatrick, C. D. (2014). The reliability of morphological analyses in language samples. *Language Testing, 31*(1), 3–18.
Tulling, M. A. (2022). Neural and Developmental Bases of Processing Language Outside the Here-and-Now [Ph.D., New York University]. In *ProQuest Dissertations and Theses*. https://www.proquest.com/docview/2708226240/abstract/877DEAEE955C4847PQ/1
Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40*(1–2), 21–81.
Van Houten, L. J. (1986). *The Role of Maternal Input in the Acquisition Process: The Communicative Strategies of Adolescent and Older Mothers with the Language*

Learning Children. Boston University Conference on Language Development (BUCLD), Boston. ERIC.
van Kleeck, A., Maxwell, M., & Gunter, C. (1985). A methodological study of illocutionary coding in adult-child interaction. *Journal of Pragmatics, 9*(5), 659–681.
von Prince, K. (2017). Counterfactuality and past. *Linguistics and Philosophy*, 1–39.
Walker, C. M., & Gopnik, A. (2013). Pretense and possibility—A theoretical proposal about the effects of pretend play on development: Comment on Lillard et al. (2013). *Psychological Bulletin, 139*(1), 40. https://doi.org/10.1037/a0030151
Warren, A. R. (1982). *Sex differences in speech to children.*
Weisberg, D. S., & Gopnik, A. (2016). Which Counterfactuals Matter? A Response to Beck. *Cognitive Science, 40*(1), 257–259. https://doi.org/10.1111/cogs.12241
Weismer, S. E., Venker, C. E., Evans, J. L., & Moyle, M. J. (2013). Fast mapping in late-talking toddlers. *Applied Psycholinguistics, 34*(1), 69–89.
Weist, R. M. (2018). Whorfian potential in child language. *Psychology of Language and Communication, 22*(1), 467–491.
Weist, R. M., Lyytinen, P., Wysocka, J., & Atanassova, M. (1997). The interaction of language and thought in children's language acquisition: A crosslinguistic study. *Journal of Child Language, 24*(1), 81–121.
Weist, R. M., & Zevenbergen, A. A. (2008). Autobiographical memory and past time reference. *Language Learning and Development, 4*(4), 291–308.
Wells, G. (1981). Language as interaction. *Learning through Interaction: The Study of Language Development*, 22–72.
Yong, Q. (2016). A corpus-based study of counterfactuals in Mandarin. *Language and Linguistics, 17*(6), 891–915.

Data, code and materials availability statement

All data, code and materials related to this study are publicly available for researchers to examine and use. If any of the links provided here become unavailable, you can request access through contacting the first author of this paper (maxime.tulling@umontreal.ca).

All used corpus data is freely available on the CHILDES Talkbank (MacWhinney, 2000): https://childes.talkbank.org/, or can be accessed through the childes-db project via R, Python or MySQL (Sanchez et al., 2019): https://langcog.github.io/childes-db-website/. The coded data, complementary datafiles and all scripts related to corpus extraction, data processing, statistical analysis and visualization are available at: https://osf.io/h2jm3/.

Authorship and Contributorship Statement

MT conceived of the study, designed the study and wrote the first draft of the manuscript. AC contributed to the design of the study and revised the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

Thanks to Ioana Grosu for intercoder reliability and Sam Mitchell, Stacy Gerchick and Mark Bacon for checking prior coding schemes. Thanks to Sudha Arunachalam and Stephanie Harves for comments and suggestions on an earlier version of this work, and thanks to Annemarie van Dooren and Yu'an Yang for their help and support.

Appendices

S1: Overview of Used Corpora

Table S1. ***Overview of all corpora used: corpus name, collection, children's age range (in months), the number of children documented, the number of utterances and wishes found separated by children and adults, and corpus citation. Shaded rows indicate corpora that did not include any*** **wish*****-utterances from children.***

Corpus	Collection	Min Age	Max Age	N	N Child Utterances (N wishes)	N Adult Utterances (N wishes)	Citation
Belfast	Eng-UK	24.1	54.2	11	25781 (1)	80899 (28)	(Henry, 1995)
Bliss	Eng-NA	40.0	64.0	4	1302 (1)	1011 (0)	(Bliss, 1988)
Bloom	Eng-NA	19.2	37.7	2	31970 (0)	36071 (NA)	(Bloom et al., 1974)
Bohannon	Eng-NA	36.0	36.0	3	4057 (0)	6737 (NA)	(Bohannon & Marquis, 1977)
Braunwald	Eng-NA	15.0	84.5	1	53311 (30)	33970 (21)	(Braunwald, 1971)
Brown	Eng-NA	27.1	62.4	2	96747 (32)	86172 (32)	(Brown, 1973)
Clark	Eng-NA	26.5	38.1	1	18185 (2)	24283 (9)	(Clark, 1979)
Compton-Pater	Eng-NA	8.0	38.7	3	25169 (1)	0 (0)	(Pater, 1997)
Cruttenden	Eng-UK	17.6	46.1	2	3061 (0)	0 (NA)	(Cruttenden, 1978)

Davis	Eng-NA	6.4	36.1	6	97128 (3)	0 (0)	(B. L. Davis & MacNeilage, 1995)
Davis-CDI	Eng-NA	8.9	35.7	4	3763 (3)	0 (0)	(Davis et al., 2018)
Demetras1	Eng-NA	24.9	47.9	1	6971 (1)	8293 (0)	(Demetras, 1989)
Demetras2	Eng-NA	26.5	33.8	1	9411 (3)	11119 (5)	(Demetras et al., 1986)
EllisWeis-mer	Clinical-MOR	30.0	66.0	13	71074 (11)	102876 (11)	(Weismer et al., 2013)
ENNI	Clinical-MOR	48.4	119.8	1	29269 (1)	650 (0)	(Schneider et al., 2006)
Evans	Eng-NA	71.3	71.3	1	4787 (0)	10 (NA)	(Evans, 1985)
Fletcher	Eng-UK	36.0	86.4	48	22073 (2)	26251 (0)	(Johnson, 1986)
Forrester	Eng-UK	12.0	60.0	1	7536 (2)	8919 (3)	(Forrester, 2002)
Garvey	Eng-NA	34.0	67.0	62	10338 (26)	9 (0)	(Garvey & Hogan, 1973)
Gathercole	Eng-NA	33.0	78.0	14	6724 (11)	2743 (1)	(Gathercole, 1986)
Gelman	Eng-NA	18.0	84.2	2	52281 (19)	126964 (32)	(Gelman et al., 1998, 2004, 2014; Jipson et al., 2016)
Gleason	Eng-NA	26.5	62.3	22	20247 (3)	38880 (6)	(Bellinger & Gleason, 1982)
Goad	Eng-NA	17.6	42.6	2	8853 (1)	0 (0)	(Parsons, 2006)
Gopnik	Eng-NA	24.0	64.7	1	3754 (1)	6347 (0)	(M. Gopnik, 1989)
Haggerty	Eng-NA	31.6	31.6	1	1739 (0)	0 (NA)	(Haggerty, 1930)
Hall	Eng-NA	54.0	57.0	36	124924 (71)	107305	(Hall & Tirre, 1979)
Hicks	Eng-NA	61.0	95.0	21	8992 (0)	5248 (NA)	(Hicks, 1991)
Higginson	Eng-NA	22.0	35.0	1	5953 (0)	9672 (NA)	(Higginson, 1985)
HSLLD	Eng-NA	42.6	141.9	11	130124 (25)	172908 (75)	(Dickinson & Tabors, 2001)
Inkelas	Eng-NA	6.3	45.9	1	1873 (0)	0 (NA)	(Inkelas & Rose, 2007)
Kuczaj	Eng-NA	28.8	60.4	1	32172 (25)	25622 (14)	(Kuczaj, 1977)
Lara	Eng-UK	21.4	40.0	1	57639 (4)	99728 (14)	(Rowland & Fletcher, 2006)
Mac-Whinney	Eng-NA	1.0	92.1	3	57675 (69)	63605 (17)	(MacWhinney, 1991)
Manchester	Eng-UK	20.7	36.3	13	249504 (5)	374198 (39)	(Theakston et al., 2001)
Morisset	Eng-NA	30.0	39.0	100	12964 (1)	19341 (0)	(Morisset et al., 1990)

MPI-EVA-Manchester	Eng-UK	24.0	37.1	2	253910 (14)	320710 (83)	(Lieven et al., 2009)
Nelson	Eng-NA	19.6	32.8	1	4552 (4)	1624 (1)	(Nelson, 1989)
New-England	Eng-NA	13.5	33.0	24	12041 (0)	43667 (NA)	(Ninio et al., 1994)
Newman Ratner	Eng-NA	11.0	288.0	1	23268 (0)	164190 (NA)	(Newman et al., 2016)
Paido-English	Eng-NA	27.0	69.0	1	10169 (0)	0 (NA)	(Edwards & Beckman, 2008)
Penney	Eng-NA	59.9	72.1	21	1491 (0)	944 (NA)	(Judd, 2018)
Peterson-McCabe	Eng-NA	48.0	113.0	1	10361 (1)	7216 (0)	(Peterson & McCabe, 1983)
Post	Eng-NA	22.7	32.2	1	16893 (0)	18755 (NA)	(Demetras et al., 1986)
Providence	Eng-NA	11.1	48.1	6	176132 (16)	283927 (109)	(Demuth et al., 2006)
Sachs	Eng-NA	15.0	57.1	1	17236 (0)	12222 (NA)	(Sachs & Nelson, 1983)
Smith	Eng-UK	26.1	45.4	1	5308 (0)	0 (NA)	(Smith, 1973)
Snow	Eng-NA	29.6	45.1	1	13520 (2)	21033 (16)	(MacWhinney & Snow, 1990)
Sprott	Eng-NA	33.0	61.0	27	4718 (2)	1606 (0)	(Sprott, 1992)
Suppes	Eng-NA	23.5	39.7	1	33950 (1)	35172 (4)	(Suppes, 1974)
Thomas	Eng-UK	24.4	59.7	2	218984 (58)	372363 (153)	(Lieven et al., 2009)
Tom-merdahl	Eng-UK	29.0	45.0	1	12027 (2)	13879 (2)	(Tommerdahl & Kilpatrick, 2014)
Valian	Eng-NA	21.7	32.8	1	15945 (1)	27715 (2)	(Valian, 1991)
VanHouten	Eng-NA	28.0	43.4	26	4455 (1)	8736 (0)	(Van Houten, 1986)
VanKleeck	Eng-NA	37.0	48.0	20	6677 (0)	8756 (NA)	(van Kleeck et al., 1985)
Warren	Eng-NA	30.0	70.0	11	3563 (0)	5847 (NA)	(Warren-Leubecker, 1982)
Weist	Eng-NA	25.0	60.2	7	47577 (8)	65165 (12)	(Weist & Zevenbergen, 2008)
Wells	Eng-UK	17.7	60.8	31	57537 (14)	40756 (11)	(Wells, 1981)
Total	NA	1.0	288.0	585	2247665 (478)	2934114 (771)	

S2: Intercoder Reliability Values

Table S2. ***Results from calculating overall accuracy (%), Gwet's AC1 coefficient and Conger's kappa statistic for each coded variable.***

Variable	Test	Value	CI (95%)
Main Subject	Percent Agreement	0.94	(0.893,0.987)
	AC1	0.94	(0.884,0.987)
	Conger's Kappa	0.80	(0.64,0.951)
Embedded Subject	Percent Agreement	0.96	(0.921,0.999)
	AC1	0.96	(0.913,0.999)
	Conger's Kappa	0.94	(0.881,0.998)
Subjunctivity	Percent Agreement	0.96	(0.921,0.999)
	AC1	0.95	(0.903,0.999)
	Conger's Kappa	0.89	(0.784,0.997)
Temporal Orientation	Percent Agreement	0.88	(0.815,0.945)
	AC1	0.87	(0.792,0.941)
	Conger's Kappa	0.60	(0.403,0.797)
Bare Error	Percent Agreement	0.93	(0.879,0.981)
	AC1	0.92	(0.871,0.982)
	Conger's Kappa	0.28	(-0.034,0.6)
Tense Error	Percent Agreement	0.89	(0.828,0.952)
	AC1	0.88	(0.807,0.95)
	Conger's Kappa	0.61	(0.439,0.79)
Evidence Counterfactuality (before discussion)	Percent Agreement	0.61	(0.513,0.707)
	AC1	0.52	(0.401,0.64)
	Conger's Kappa	0.49	(0.358,0.612)
Evidence Counterfactuality (after discussion)	Percent Agreement	0.87	(0.803,0.937)
	AC1	0.84	(0.757,0.922)
	Conger's Kappa	0.83	(0.743,0.918)

S3: Supplement to Figure 3

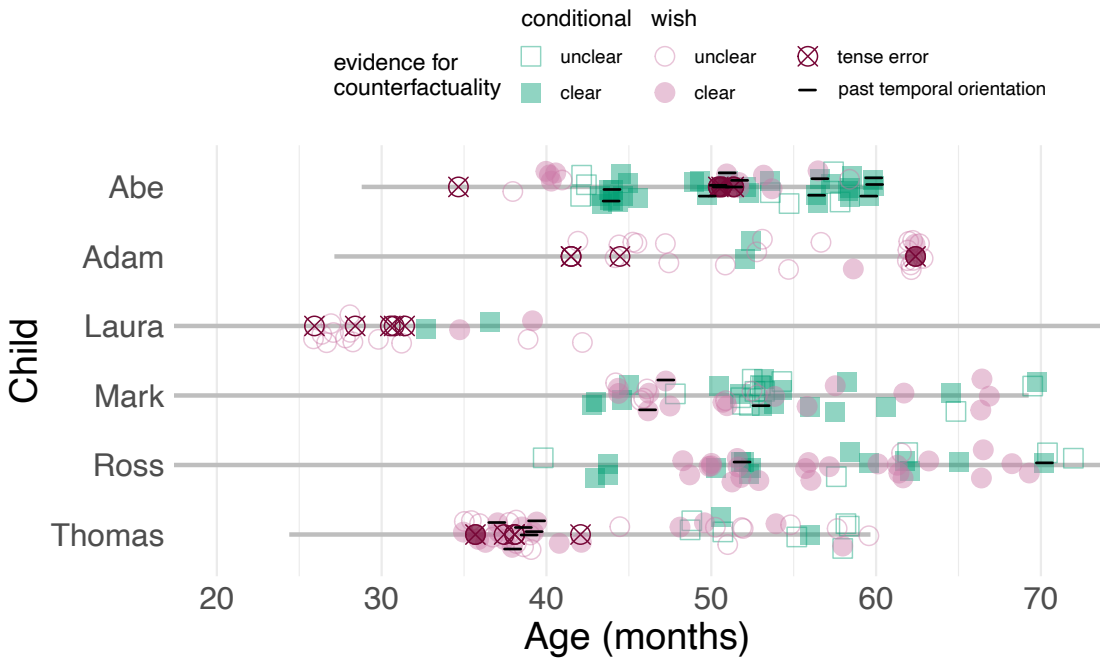


Figure S3. ***Counterfactual conditionals (green squares) and wishes (pink circles) for each child (y-axis) with age indicated in months on the x-axis. Filled shapes indicate that the evidence for counterfactuality is clear, empty shapes indicates the evidence is unclear. Struck-through (with cross) wishes indicate they contained a tense error in the form of a present-for-past substitution. Struck-through (with black dash) counterfactuals were uttered with past temporal orientation, all others are present temporal orientation. Grey line indicates recording span.***

S4: Overview of Children's Productivity with the Past Tense

Table S4. ***Overview of Children's Past Tense Productivity. For each child we recorded their age range (in months), total amount of utterances, total amount of produced present-for-past errors, age range while making errors, the proportion of correct past tense marking in the context of the temporal adverb yesterday (YD), total amount of past tense overregularization (OR) and age range of during which overregularized.***

Child	Abe	Adam	Laura	Mark	Ross	Thomas
Corpus	Kuczaj	Brown	Braunwald	Mac-Whinney	Mac-Whinney	Thomas
Min Age	28.8	27.1	15.0	5.5	16.4	24.4
Max Age	60.4	62.4	84.5	69.3	92.1	59.7
N Utterances	31958	46651	39750	20754	36379	218439
N Errors	4.0	4.0	5.0	*NA*	*NA*	5.0
Error Min Age	34.7	41.5	25.9	*NA*	*NA*	35.7
Error Max Age	51.4	62.4	31.4	*NA*	*NA*	42.1
N Past with YD	13/14	3/3	2/6	*NA*	*NA*	*NA*
YD Min Age	34.7	55.0	28.0	*NA*	*NA*	*NA*
N OR	218.0	22.0	8.0	*NA*	*NA*	22.0
OR Min Age	34.7	42.3	26.2	*NA*	*NA*	35.9
OR Max Age	51.2	62.4	31.0	*NA*	*NA*	42.1

License