Caregiver-infant interactions and child vocabulary

A large-scale, longitudinal study of dyadic and multimodal behaviours

Published by
LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: Emma de Beer

ISBN: 978-94-6093-449-0
DOI: https://dx.medra.org/10.48273/LOT0664
NUR: 616

Caregiver-infant interactions and child vocabulary

A large-scale, longitudinal study of dyadic and multimodal behaviours

Ouder-kind interacties en de woordenschat van kinderen

Een grootschalig, longitudinaal onderzoek naar dyadische en multimodale gedragingen (met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 8 maart 2024 des ochtends te 10.15 uur

door

Anika van der Klis

geboren op 12 oktober 1994
te Woerden

Promotor:

Prof. dr. R.W.J. Kager

Copromotor:

Dr. F.W. Adriaans

Beoordelingscommissie:

Prof. dr. C.C. Levelt
Prof. dr. S. Peperkamp
Dr. J. Verhagen
Prof. dr. G. Vigliocco
Prof. dr. F.N.K. Wijnen (voorzitter)

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van het Zwaartekrachtprogramma van het Nederlandse Ministerie van Onderwijs, Cultuur en Wetenschap en de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO-subsidienummer 024.001.003).

Table of Contents

List of Tables

List of Figures

Acknowledgements

Although this is the first section of the dissertation, writing this marks the end of my PhD journey. I would like to thank several people for supporting me over the past four years and making this dissertation possible.

First and foremost, I would like to thank my supervisors René Kager and Frans Adriaans. Throughout the project, we have met regularly – in Utrecht or virtually – despite a global pandemic. Your kindness, clear (and fast!) feedback, guidance, and warm encouragements were very motivating. I would always feel more relaxed after our meetings, no matter how stressed I felt beforehand. I am also very happy with the amount of freedom you gave me to explore different areas of research. This have made the articles in this dissertation diverse and interdisciplinary. You have truly been very supportive throughout the whole process, and I could not have wished for better supervisors.

I would also like to specially thank Caroline Junge. Although not officially my PhD supervisor, your contributions to the project as a second author on two of the chapters in my dissertation have been very valuable. Thank you for your collaboration, and for teaching me the ins and outs of collecting and interpreting data on children's receptive and expressive vocabulary development.

I would also like to thank all members of the assessment committee: Claartje Levelt, Sharon Peperkamp, Josje Verhagen, Gabriella Vigliocco, and Frank Wijnen. Thank you for reading this dissertation and providing me with useful comments and suggestions.

Then, I wish to thank my paranymphs: Zsofia Bélteki and Rachida Ganga. We started our PhD projects around the same time, and it feels only right to finish with you standing beside me during the defence. I am happy that we could be each other's

support during good and bad times. Thank you for supporting me on this journey one more time.

I would also like to thank my office mates throughout the years: Quy Ngoc Thi Doan, Mo Chou, Mariano Gonzalez, Kexin Du, and Victoria Reshetnikova. We have not seen each other as regularly as we would have without the pandemic, but it was fun sharing an office (and the occasional drink) with you. I also would like to give a shoutout to my office neighbours: Emma Everaert, Joanna Wall, and Sofiya Ros. Having to cross your office in order to enter my own does create a bond. And of course, thanks to all other ILS PhD students with whom I have chatted in the hallways, frequently had lunches, or enjoyed the Uiltjesdagen. You all contributed to the enjoyable time I had as an ILS PhD student.

I also wish to warmly thank Maaike Schoorlemmer. You make sure all new ILS PhD students feel welcomed and quickly at home in the institute. Thank you for all your help and support.

I would also like all the researchers involved in the Babylab in Utrecht. A special thanks is due to Desiree Capel and Frank Wijnen for organising the weekly lab meetings. While I did not test infants in the Utrecht Babylab for my projects, the insights and discussions held during the lab meetings were invaluable to my research. I would also like to thank frequent attendees, including Rachida Ganga, Iris van der Wulp, Areti Kotsolakou, Caroline Junge, Sita ter Haar, Karin Wanrooij, Charlotte Koevoets, Victoria Reshetnikova, Shalom Zuckerman, Hugo Schnack, Marijn Struiksma, Elanie van Niekerk, Jorik Geutjes, and Aoju Chen. Thank you for the insightful but also fun meetings every Friday morning in Utrecht!

Then, I wish to express a special thanks to all people that were involved in the YOUth cohort study and the KinderKennisCentrum in Utrecht. First, I wish to thank Chantal Kemner. Because of your efforts over the past ten years, CID and the YOUth cohort study have resulted in invaluable longitudinal data set on child development. I would

also like to mention Femke Everaarts, Gwen Schouw, Ivonne Fokkert, Lilli van Wielink, Liset de Gee, Mark Bruurmijn, Leon Versteegen, and Jorinde Pasman for making every day at the KKC very enjoyable. And of course, a special thanks to all research assistants and researchers involved in YOUth. Last but not least, I wish to thank Ron Scholten for being a wonderful data manager. Thank you for your incredibly swift replies to all my (many) questions.

Additional thanks go to my family members and close friends. Thank you for always listening and responding to all my ideas, doubts, enthusiasm, complaints, laughter, and cries at every hour of the day. You are a safe space for me to share my thoughts and vent my feelings which have kept me sane.

Utrecht,
January 2024

Chapter 1

Introduction

Infants are born in complex environments full of auditory and visual signals from which they have to discover rules and meanings. To communicate effectively, children must learn the names of objects, actions, or events. Infants already start learning their native language in the womb. At birth, infants show behavioural and neural evidence of being able to discriminate their native language – which they have heard in utero – from another language (e.g., May et al., 2011; Mehler et al., 1988; Moon et al., 2013). At this point, infants are not yet able to segment native words from continuous speech (see Jusczyk, 1999). By six to nine months of age, however, infants already know the meanings of some common nouns, such as *hands* or *banana* (Bergelson & Swingley, 2012). After segmenting words from continuous speech, infants have to learn the meanings of those words. This entails a computational problem: hearing a novel word and seeing a scene presents the learner with an infinite number of possible referents. This referential ambiguity is famously referred to as the "gavagai" problem. Quine (1964) uses the example that hearing a speaker of an unknown language say "gavagai" when seeing a rabbit, the listener would assume it translates to rabbit. But how do we know that the speaker is not referring to a smaller part of the rabbit, such as "furry", or the activity that the rabbit is performing, like "hopping"? Infants learning their native language(s) have to solve such referential ambiguities in the language input they receive from their caregivers in order to learn word meanings and develop their vocabularies. Previous studies have examined what information infants can use to do this. For example, research has suggested that infants may be able to use social cues, such as pointing gestures and body orientation (e.g., Baldwin et al., 1996; Grassmann & Tomasello, 2010). Such cues embedded in social interactions could help infants resolve referential ambiguities and subsequently learn words. This dissertation investigates social interactions between infants and caregivers and children's vocabulary outcomes.

2 Caregiver-infant interactions and child vocabulary

Traditionally, theories of social learning have emphasised the importance of social interaction for language learning (Bruner, 1983; Vygotsky, 1962). To date, many studies have found evidence for the impact of social factors on children's language development (for reviews, see Hoff, 2006; Kuhl, 2007; Rowe & Weisleder, 2020). Social interactions are marked by contingency. For example, if an infant reaches for a rattler and the caregiver hands it over while saying "Here is your rattler", the caregiver's response is contingent because it is quick and appropriate to the infant's behaviour (e.g., Skinner, 1986; Tamis-LeMonda et al., 2014). Studies propose four means through which caregivers' contingent responses could facilitate learning: temporal, semantic, pragmatic, and attentional (Kuhl, 2007; Masek et al., 2021a; Tamis-LeMonda et al., 2014). First, the temporal contingency of a caregiver's responses to an infant's behaviour – when the response occurs shortly after the behaviour – makes it easier for the infant to understand that the two events are linked (Jaffe et al., 2001; Keller et al., 1999). Second, the semantic contingency – when the verbal contents of the response are related to the object or activity that the infant is paying attention to – could make it easier for the infant to link the contents of the caregiver's response to the environment (Baldwin & Markman, 1989; Carpenter et al., 1998). Third, contingent responses could have a pragmatic function. Infants learn that they can communicate effectively by producing sounds and gestures, and they learn the functions of different types of gestures, such as the deictic (i.e., to refer to an object or person) or requesting (i.e., to obtain an object out of reach) functions of pointing or reaching gestures, respectively (Blake et al., 1994). They can then use these communicative acts to establish joint attention, obtain information, and learn from adults (see Tamis-LeMonda et al., 2014). Lastly, the relationship between contingent interactions and language learning could be driven by the infant's increased attention to the caregiver's response (Chen et al., 2021; Kuhl., 2007; Masek et al., 2021a). Despite more than half a century of research in this area, many areas remain unexplored due to the large complexity of analysing naturalistic caregiver-infant interactions.

First, studies have largely focused on caregivers' verbal responsiveness, while communication is **multimodal**. Caregivers' temporally and semantically contingent verbal responses to infants' vocalisation and gestures have been found to predict children's vocabulary outcomes (e.g., Bornstein et al., 2008; McGillion et al., 2013; Olson & Masur, 2015; Tamis-LeMonda et al., 2001; Wu & Gros-Louis, 2015). The temporal and semantic contingency of the caregiver's response could make it easier for the infant to map the phonological form of the word to objects or events, reducing the referential ambiguity in the language input. Yet, communication is inherently multimodal with audio and visual information often being provided concurrently. When an infant points at a doll, and the caregiver says: "What a pretty doll!", while showing it to the infant, the infant receives two simultaneous cues for the name of the object from the caregiver: an audio and a visual cue. The combination of an audio and visual cue makes the input less ambiguous which could make it easier for the infant to map the word "doll" onto its meaning referent (e.g., Baldwin et al., 1996; Gogate et al., 2000). Yet, studies examining the relationship between caregivers' responsiveness and children's vocabulary outcomes rarely take caregivers' nonverbal or multimodal (i.e., coordinated verbal and nonverbal) responsiveness into account. A caregiver's multimodal response potentially minimises referential ambiguity in the learning context – making this an excellent word-learning opportunity for the infant.

Second, many studies have focused on either the infant *or* the caregiver and how their individual behaviours during interactions relate to children's vocabulary outcomes, while communication is **bidirectional**. For example, many studies examined the association between infants' pointing gestures and their vocabulary outcomes without taking into account any of the caregivers' behaviours (see Colonnesi et al., 2010) or caregivers' quality and quantity of language input and their children's vocabulary development without taking into account any infants' behaviours (see Anderson et al., 2021). Crucially, learning occurs in shared interactions that are constructed by both the infant and the caregiver (see Renzi et al., 2017). Therefore, the question remains how children and caregivers jointly contribute to the word learning experience. Chen et al. (2021) recently found that caregivers' naming of objects tended to follow

children's attention, measured by their looking behaviour, when the object was unfamiliar to the child. When the object was unfamiliar, caregivers also tended to touch the object more often while labelling it compared to a familiar object. Furthermore, when caregivers touched the object while labelling it, this resulted in extended looking behaviour (i.e., visual attention) by the child (Chen et al., 2021). This example shows how the infant's and caregiver's individual behaviours affect the other's behaviours during real-time interactions. In another study, Ger et al. (2018) found that some infant behaviours, such as hand shape or vocalising during pointing, influenced caregivers' responsiveness to infants' pointing gestures. The study also found that caregivers' semantically contingent responses to infants' pointing gestures at 10 months related to an increase in infants' pointing at 12 months. Analysing the joint behaviours of caregivers and infants helps us to understand how both contribute to the learning environment, which could improve our understanding of how children successfully learn words.

The overarching goal of this dissertation is to predict variation in Dutch children's vocabulary skills using data from the large-scale, longitudinal YOUth cohort study. We take a dyadic approach to study the effects of verbal, nonverbal, and multimodal behaviours during caregiver-infant interactions. To achieve this, we need to annotate verbal and nonverbal behaviours in caregiver-infant interactions and have reliable measures of Dutch children's vocabulary outcomes. This dissertation consists of four empirical articles that address methodological and theoretical gaps relating to the overarching goal (Chapters 2–5). This introduction provides a review of the literature, addresses methodological considerations, identifies the research gaps, and introduces the research questions driving the four studies.

1.1. Vocabulary and gesture development

Around their first birthdays, infants typically achieve a milestone: The production of their first word. The first words of infants across the world are remarkably similar. They typically involve the word for "mother" or "father" (e.g., *mommy*, *daddy*) quickly followed by social routines (e.g., *bye*, *yum yum*), animals and animal sounds

(e.g., *woof woof*, *cat*), and foods (e.g., *bread*, *banana*) (Frank et al., 2021; Tardif et al., 2008). While it takes infants some time to start producing their first words, the rate of vocabulary development speeds up linearly across the second year (Frank et al., 2021). Studies suggest that children undergo a vocabulary "spurt", suggesting that children's vocabulary developmental rate increases over time (Ganger & Brent, 2004; cf. McMurray, 2007). Different processes, such as syntactic bootstrapping (i.e., learning words by recognising syntactic categories) or mutual exclusivity (i.e., a constraint which guides children to map unknown words to unfamiliar rather than familiar referents) make it gradually easier to learn new words by using knowledge of known words (e.g., Markman, 1991; Naigles & Swensen, 2007; Nazzi & Bertoncini, 2003). Despite the remarkable similarity across languages in children's first words and the acceleration in their developmental rates, children's early development is also marked by large individual differences in the onset of words and gestures and total vocabulary size.

The sizes of infants' vocabularies and gesture repertoires are often measured through caregiver reports, such as the MacArthur–Bates Communicative Development Inventories (CDIs) (Fenson et al., 2007). Caregivers receive a list of vocabulary items, and they mark for each item whether their child understands and/or speaks the word. For American English 24-month-olds, many children are reported to have small vocabularies (i.e., zero up to a hundred items on the checklist), while others are reported to already produce all vocabulary items included on the checklist (i.e., over six hundred words) (Frank et al., 2021). These caregiver reports were one of the first reliable tools that could capture variation in infants' vocabulary on a very large scale. The results of the first norming study revealed the large individual variability in children's vocabulary that is already present before their first birthdays (Fenson et al., 1994) which has captivated the interest of researchers in this field to date. The large differences found for infants tend to remain stable throughout childhood (Bornstein & Putnick., 2012; Fenson et al., 1994) and have lasting effects on children's later language, reading, and other social and academic skills (Bleses et al., 2016; Morgan et al., 2015; Preston et al., 2010). Early talkers remain at an advantage compared to

late talkers which makes it relevant to study predictors of variation in children's early vocabulary development.

Children start communicating long before they produce their first words. From birth, infants start crying and cooing to grab their caregivers' attention. They can show facial expressions like smiling and grimacing to indicate how they feel. By five months of age, infants have learned that their vocalisations influence the behaviour of others (Goldstein et al., 2009). At nine months, infants restructure their own vocalisations to match the phonological patterns that they hear in verbal responses from their caregivers (Goldstein & Schwade, 2008). Starting from at least ten and twelve months of age, infants use showing and pointing gestures respectively with the goal to share attention and interest with others (Boundy et al., 2019; Liszkowski et al., 2004). Children's gestures are precursors to and predictors of their later vocabularies. Infants that produce gestures earlier, produce more types of gestures, or produce higher gesture rates tend to have larger vocabulary sizes later in life (e.g., Brooks & Meltzoff, 2008; Colonnesi et al., 2010; Rowe & Goldin-Meadow, 2009). One hypothesis is that by producing gestures, infants create word-learning opportunities for themselves. Caregivers tend to respond to gestures, particularly to pointing gestures, by providing timely object labels or other relevant semantic information (Olson & Masur, 2011, 2013; Wu & Gros-Louis, 2015). For example, when an infant points towards a cat, the caregiver is likely to respond with "That's a cat!". The more gestures infants produce, the more labelling responses they tend to elicit from their caregivers. A previous study also showed that such labelling responses mediate the relationship between infants' gestures and their vocabulary outcomes (Olson & Masur, 2015). This suggests that infants' gestures relate to their vocabularies because they tend to elicit verbally contingent responses from their caregivers.

There are challenges when collecting data on children's vocabularies and gesture inventories. For infants, the most common method is collecting caregiver reports. Using caregiver reports is a cost-effective approach because they do not require trained lab assistants or lab visits. This is in stark contrast with naturalistic speech

samples which require lab or home visits and the labour-intensive transcription of speech data. Yet, caregiver reports are prone to caregiver reporting biases, such as overreporting based on societal expectations which makes caregivers believe larger vocabularies are more desirable, or underreporting because caregivers have different criteria for determining word comprehension (i.e., whether a child understands a word) (for discussions on this, see Feldman et al., 2000; Tomasello & Mervis, 1994). Naturalistic speech samples have high ecological validity and do not require any data interpretations by the caregiver. Yet, recorded speech samples are limited snapshots which mostly reflect which words or linguistic structures a child uses, while caregiver reports indicate what a child knows. It is also more difficult to standardise speech samples and avoid other confounds (discussed in Frank et al., 2021). For toddlers, it is possible to collect vocabulary data using lab-administered tasks, such as the Peabody Picture Vocabulary Task (PPVT; Dunn & Dunn, 1997). Although these could be less influenced by caregivers or the limited durations of speech recordings, toddlers can be unwilling to cooperate during lab-administered tasks which negatively influences their performance. To summarise, there are different methods to collect data on children's vocabularies, each with its own set of difficulties.

For children growing up in the Netherlands learning Dutch, there are several standardised vocabulary tasks available. In this dissertation, we used the Dutch adaptation of the Peabody Picture Vocabulary Task for toddlers (PPVT-III-NL; Schlichting, 2005). This standardised task measures receptive vocabulary by counting the number of spoken words that participants can match to one of four pictures. Words become increasingly more complex. If participants make too many errors (> 8) within one set of twelve items, the task terminates. For each participant, this results in a raw score (i.e., total number of correct items) which corresponds to a normed score based on the child's age. This task was administered in the lab. Then, we also used caregiver reports of children's vocabularies during infancy and toddlerhood. We adapted short forms of Dutch versions of the CDIs (N-CDIs; Zink & Lejaegere, 2002, 2003). The N-CDIs were originally designed and normed for children growing up in Belgium speaking Flemish Dutch. To better match the ages and language of children

participating in the YOUth cohort study, the N-CDIs had to be adapted in several ways. First, we used short forms to save time as caregivers in the YOUth cohort study have to fill out a broad range of questionnaires. We included the gesture scale from the full-length N-CDI-Words and Gestures which is not normally included when administering short forms, while this could be a more relevant scale for capturing individual variation across infants aged 9–11 months (Zink & Lejaegere, 2002). Second, we changed or removed Flemish Dutch words to better fit the dialect of Dutch spoken in the Netherlands. Lastly, we combined the N-CDI 2 and N-CDI 3 to better accommodate the age range of children included in the second measurement wave. The adaptations make the N_{YOUth}-CDIs more suitable for the participants included in the YOUth cohort study.

In short, normally developing infants tend to acquire communicative gestures and words quite rapidly throughout the first few years of life. Despite the consistency in the types of words and gestures that are acquired first, infants show large variability in the onset and frequency with which they produce vocalisations and gestures (Frank et al., 2021). These early differences have predictive value for children's later language skills and other cognitive outcomes which makes it relevant to 1) identify which vocabulary tasks are most successful at capturing individual variation across children in different age groups, and 2) identify factors which can predict some of the large variability.

1.2. Influences of verbal and nonverbal input

Environmental factors most significantly influence variation in children's vocabularies in the early years (see Kidd & Donnelly, 2020). To successfully learn a language, infants must receive language input to learn from. Research suggests that caregivers vary widely in the language input they provide to their children. Although researchers have long focused on the quantity of spoken language input, the contents of speech and the way that caregivers communicate with their children have a larger influence on children's linguistic outcomes (see Anderson et al., 2021; Masek et al., 2021b). Caregivers of higher socio-economic status tend to provide more speech (i.e.,

higher quantity) and more diverse speech (i.e., higher quality) which positively affects their children's vocabulary development (e.g., Hoff, 2003; Huttenlocher et al., 2010; Rowe, 2012). Hart and Risley (1995) recorded the number of words that children hear per waking hour. They found such a large difference between the average number of words addressed to children from high-income families and low-income families, that in four years, this difference would add up to a "30-million-word gap" in the language input (cf. Sperry et al., 2019). Speech quantity and speech quality show differential effects on children's vocabularies at different developmental stages (Rowe, 2012). Of all forms of language input, immediate verbal responses to infants' vocalisations and gestures provide unique contributions to children's vocabulary development (e.g., Bornstein et al., 1999; Hoff, 2003; McGillion et al., 2013). McGillion et al. (2013) found a relation between caregivers' responsiveness and children's linguistic outcomes after controlling for caregivers' quantity of speech overall. These studies show the importance of caregivers' language input for shaping children's early learning opportunities.

Besides the quantity and quality of caregivers' speech, caregivers also differ in the degree to which they modify the acoustic signal of their speech addressed to infants. Infant-directed speech (IDS), "baby talk" or "motherese" refers to the spontaneous prosodic modifications by adults, such as a higher mean pitch, a larger pitch range, and greater pitch variability, when addressing infants (for a review, see Soderstrom, 2007). These prosodic modifications have been positively related to children's vocabulary growth (for a meta-analysis, see Spinelli et al., 2017). It remains debated how these acoustic properties facilitate children's vocabularies. Studies have shown that slow speech improves children's word recognition abilities (e.g., Song et al., 2010; Zangl et al., 2005). The results for the pitch of IDS are less conclusive. Experimental studies have shown that infants only learn novel words when they are presented in IDS but not in adult-directed speech (ADS) (Estes & Hurley, 2013; Singh et al., 2009), but the effects of pitch have not been studied in isolation from the other acoustic properties of IDS. Some studies suggest that the prosody of IDS facilitates infants' ability to segment words from continuous speech by highlighting structures

(Jusczyk et al., 1992; Soderstrom et al., 2008), but Estes and Hurley (2013) showed that IDS can promote word learning even when it does not serve to function as a cue for word segmentation. Apart from direct linguistic effects, there may be indirect effects, such as that children appear to prefer to listen to IDS versus ADS (Fernald, 1985; Dunst et al., 2012; Soderstrom, 2019) which could also enhance their learning abilities. In short, the degree to which caregivers adjust the acoustics of their speech signal when addressing infants directly or indirectly influences children's vocabulary development.

Apart from spoken language, caregivers also use a range of nonverbal behaviours, such as gestures or touch, to communicate with their infants. Previous studies found that children use gaze direction, body orientation, and index-finger pointing as cues to learn the reference of novel words (e.g., Baldwin et al., 1996; Grassmann & Tomasello, 2010). Caregivers' actions directed at infants are less complex, closer in proximity, have a larger range of motion, and contain more repetitions compared to actions directed at other adults, which has been referred to as "motionese" (Brand et al., 2002). Brand et al. (2002) hypothesised that these modifications in caregivers' nonverbal communication have a similar function as the modifications in the acoustic signal of speech addressed to infants: To maintain infants' attention and to highlight structures and meaning. While letting caregivers teach new words to their infants during recorded play sessions, Gogate et al. (2000) found that most maternal utterances are multimodal – involving audio and visual information including gestures or touch – and 60% of utterances are temporally synchronous with object motion. Similarly, Vigliocco et al. (2019) found that approximately 40% of all maternal utterances during caregiver-toddler interactions with a set of toys were accompanied by iconic gestures, representational gestures, or other hand actions, such as deictic gestures or depicting actions (e.g., demonstrating the use of a toy). Infants can use synchronised information from an early age. For example, intersensory redundancy, such as the simultaneous showing and naming of an object, helps preverbal infants to map vowel sounds onto objects (Gogate & Bahrick, 1998). Therefore, nonverbal

behaviours could facilitate children's vocabulary development by both reducing referential ambiguity and increasing infants' attention.

Verbal language input (e.g., quantity, quality, contents, acoustics) is typically studied through audio or video recordings, while nonverbal behaviours (e.g., gestures, facial expressions) are studied through video observations. Caregivers and their infants typically visit the lab where the audio or video recording takes place in a controlled environment, or researchers visit people's homes for more ecologically valid data. To convert audio and video recordings into data which can be analysed, researchers have to carry out the time-consuming and labour-intensive task of data transcription and annotation. Typically, this involves the manual transcription of speech data and/or the manual annotation of nonverbal behaviours, such as predefined communicative gestures. Segmenting, annotating, and transcribing an hour of speech can take up to fifty hours in total, including the verification of quality (Barras et al., 2001). Annotating the content of videos frame by frame takes even longer, depending on the number of different nonverbal behaviours being analysed. Therefore, the annotation of audio and video recordings is a challenging, lengthy, and expensive task. The difficulties with the manual annotation process can also lead to smaller sample sizes which can be problematic if there is not enough statistical power (see Oakes, 2017). In the YOUth cohort study, caregiver-infant interactions are recorded in the lab. Although lab observations are more artificial than home observations, the clear advantage is that this is a controlled environment where all dyads are recorded in the same room using the same standard set of toys. This makes the data across dyads highly comparable.

1.3. YOUth cohort study

The data presented in this dissertation are derived from the longitudinal YOUth cohort study run at Utrecht University and the University Medical Centre Utrecht (see Onland-Moret et al., 2020) which is part of the Consortium on Individual Development (CID). The consortium investigates the interplay of child characteristics and environmental factors and how these inform and predict individual differences in

children's social competence and self-regulation. Social competence includes communicative competence which reflects children's knowledge and ability to use language appropriately and effectively (Fabes et al., 2006; Hymes, 1972). Communication skills are important foundational skills underlying social competence (e.g., Rose-Krasnor, 1997). Children with better language abilities tend to have more social competence, which reduces the risk of behavioural and emotional problems and increases their functioning in society (for a review, see Junge et al., 2020). The research conducted for this dissertation is a subproject of CID which focuses on children's linguistic outcomes. More specifically, children's productive and receptive vocabulary size. Children's vocabulary sizes significantly predict their subsequent language and literacy achievement (e.g., Bleses et al., 2016; Lee, 2011; Morgan et al., 2015) making vocabulary size a good early indicator of children's later language skills. Therefore, the findings presented in this dissertation inform the overarching goal of CID: how child characteristics and environmental factors result in individual differences in the development of social competence.

YOUth stands for "Youth Of Utrecht". The cohort study follows ca. 2,500 infants from the womb into childhood. Data collection started in 2013 and finished in 2023. Throughout this period, all pregnant women in the province of Utrecht in the Netherlands could start participating in the cohort study. The province of Utrecht is a densely populated region that contains both urban and rural areas. Children were only excluded from participating if they were mentally or physically incapable of completing the tasks during the lab visit. All participants were required to understand Dutch for all the information, instructions, and questionnaires (Onland-Moret et al., 2020). There were two measurement waves during infancy ("Around 0"): When the infant was around 5 months of age and around 10 months of age. Then, there was a follow-up wave when the infant was 2–4 years of age ("Around 3"). During all measurement waves, we collected videos of caregiver-child interactions. When the infant was around 10 months of age, we collected the first caregiver report on children's word production, word comprehension, and gesture repertoires (N_{YOUth}-CDI 1) which we used to determine infants' concurrent vocabulary outcomes. When

the infant was around 3 years of age, we collected another caregiver report on children's word production (N_{YOUth}-CDI 2). We also administered the PPVT-III-NL measuring children's receptive vocabularies during this wave. We used these measures to predict children's longitudinal vocabulary outcomes in this dissertation. By using a lab-administered task alongside caregiver reports, we minimise the chances that the interpretations of our results are affected by caregiver reporting biases in the data. During both measurement waves, caregivers also filled out a range of other questionnaires which we use to determine caregivers' socio-economic status, children's gestational age and birthweight, and the languages spoken at home for the studies included in this dissertation.

The YOUth cohort study results in a large dataset of children including several vocabulary outcome measures across children's development. This makes the YOUth cohort study an ideal dataset to answer questions on the large variability in children's vocabulary development. We can address questions such as the onset and stability of the effects of different predictors of variation in children's vocabulary over time. In addition, we can study whether the effects of these factors are consistent in terms of their strength and direction across different vocabulary outcomes, such as vocabulary production, vocabulary comprehension, or gesture repertoires, and across different measurement methods; namely, caregiver reports and the lab-administered task. Lastly, we can assess whether behavioural measures of early infant-caregiver interactions can predict variation in children's concurrent (i.e., at the time the infant-caregiver interactions were observed) and longitudinal (i.e., at a later point in time) vocabulary outcomes. Yet, such a large dataset presents a clear challenge. It is time-consuming and labour-intensive to annotate all data. In the YOUth cohort study, we have used digitised and automated versions of the caregiver reports and PPVT-III-NL receptive vocabulary task which drastically speeds up data processing. Therefore, this challenge is relevant for analysing the video data obtained during the caregiver-child interaction task.

1.4. Research gaps

In this section, we identify five methodological and theoretical research gaps from the literature discussed above. The first gap concerns a methodological issue regarding data annotation. The labour-intensive process of manually annotating speech data can become an issue when working with large datasets. While existing automated tools could speed up the manual annotation process, their effectiveness for the annotation of IDS is currently unknown. A growing number of automatic speech recognition (ASR) tools developed for ADS are available cross-linguistically, but IDS has specific acoustic properties that might pose a challenge for tools developed for ADS. Using a large sample from a cohort study raises the question of whether we can improve the time that is currently needed for the manual annotation process. This would allow smaller-scale research projects to include more data collected within cohort studies. The first gap related to data annotation is as follows:

> **Gap 1: The effectiveness of existing ASR tools for the automatic annotation of Dutch IDS is currently unknown.**

The second gap concerns a methodological issue regarding the validity of caregiver reports of Dutch children's vocabularies. There are currently no normed or validated caregiver reports of infants' and toddlers' vocabulary for Dutch children growing up in the Netherlands specifically. The N-CDIs were developed for Flemish Dutch (Zink & Lejaegere, 2002, 2003). Adaptations of CDIs in other languages have two advantages: they allow for testing similarities in cross-linguistic patterns (e.g., Frank et al., 2021) as well as offering a unified tool suitable for a wide range of languages to examine individual variation in vocabulary development (e.g., Cristia et al., 2014). To make the N-CDIs more suitable for the participants included in the YOUth cohort study, they were adapted in several ways – including removing or replacing Dutch Flemish items to better fit the dialect of Dutch spoken in the Netherlands. These adaptations require us to examine the validity and reliability of the N_{YOUth}-CDIs before

we continue to use them to study individual variation in Dutch children's vocabularies. The second gap related to caregiver reports of vocabulary is as follows:

> **Gap 2: The adapted N-CDIs used in the YOUth cohort study need to be examined for their validity and reliability before we continue to use them to study variation in children's vocabularies.**

The third gap is a knowledge gap concerning the onset and stability of well-known demographic predictors of variation in children's vocabularies. A limited number of studies have examined demographic predictors, such as children's gender and maternal education, on vocabulary outcomes from infancy to toddlerhood within one large, longitudinal sample. Previous studies often report different findings regarding such predictors. For example, girls outperform boys on many vocabulary scales included in the CDIs (Frank et al., 2021; Reese & Read, 2000; Zink & Lejaegere, 2002, cf. Bavin et al., 2008). In contrast, previous studies using naturalistic speech samples or lab-administered tasks of children's receptive vocabulary often do not find gender differences (Huttenlocher et al., 2010; Pan et al., 2004), although these findings are inconsistent (Bornstein & Haynes., 1998; Frank et al., 2021). Similarly, some studies report positive effects of maternal education on toddlers' vocabularies (Feldman et al., 2000; Fenson et al., 2007, cf. Reese & Read, 2000), while some studies report negative effects of maternal education on infants' vocabularies (Bavin et al., 2008; Feldman et al., 2000; Reese & Read, 2000). The uncertainty regarding the onset and stability of well-known demographic predictors of variation raises the question of whether the effects are age-specific or task-specific. Discrepancies in research findings between studies could be caused by differences in sample characteristics, task characteristics, or different developmental stages of children included in the studies. If they are caused by task characteristics, these differences question the validity of the vocabulary tasks. The third gap concerning a theoretical issue relating to the onset and stability of demographic predictors is:

Gap 3: We do not know whether well-known demographic predictors of variation in Dutch children's vocabularies are age-specific or task-specific.

The fourth gap concerns a theoretical gap related to dyadic behaviours during caregiver-infant interactions. The influence of infants' behaviours or caregivers' responses on children's vocabulary is often studied from the perspective of the infant *or* the caregiver. Typically, previous studies either examined the influence of infants' vocalisations and gestures or the influence of caregivers' language input on children's vocabulary outcomes. Yet, word learning during interactions requires dyadic behaviours from the infant and their caregiver to optimise learning (see Renzi et al., 2017). The contributions of only the child or the caregiver may not be sufficient to explain how certain behaviours during caregiver-infant interactions relate to children's vocabularies over time, especially because the behaviours of the child are dependent on the behaviours of the caregiver and vice versa. Previous studies suggest that infants' gestures are significantly related to their vocabulary development because they tend to elicit contingent verbal responses from caregivers (Olson & Masur, 2015). Infants' behaviours that elicited contingent responses from caregivers are better predictors of children's vocabulary than infants' behaviours regardless of caregivers' responses, suggesting that learning is optimised when caregivers respond to infant behaviours with appropriate language (Donnellan et al., 2019). The influences of different dyadic combinations of infants' behaviours and caregivers' responses on children's vocabulary outcomes have not been studied systematically. The fourth gap concerns the unknown relationship between dyadic behaviours and children's vocabulary outcomes:

Gap 4: The influences of different dyadic combinations of infants' vocalisations and gestures and caregivers' responses on children's vocabulary outcomes is currently unknown.

The last gap also concerns a theoretical issue related to caregivers' responsiveness. Previous studies on caregivers' responsiveness have predominantly focused on verbal language (Bornstein et al., 2008; McGillion et al., 2013; Olson & Masur, 2015; Tamis-LeMonda et al., 2001; Wu & Gros-Louis, 2015). Yet caregivers' communication with infants is multimodal (Gogate et al., 2000; Vigliocco et al., 2019). Previous studies have shown that language input contains overlapping information in the verbal and nonverbal domains that could facilitate children's vocabulary development (Gogate & Bahrick, 1998; Gogate et al., 2000). Studying language input from a multimodal perspective 1) gives us more ecologically valid data on dyadic behaviours during early caregiver-infant interactions which better describe infants' learning environments, and 2) allows us to analyse the contributions of such multimodal and dyadic behaviours in explaining variation in children's vocabulary outcomes. This informs us which specific aspects of infants' learning environments contribute to their vocabulary development, which can inform theory and intervention studies. The fifth gap is:

> **Gap 5: The influence of caregivers' multimodal responses on children's vocabulary outcomes is currently unknown.**

The overarching goal of this dissertation is to predict variation in Dutch children's vocabulary outcomes using data from the large-scale, longitudinal YOUth cohort study. We conducted four empirical studies to address the research gaps. Addressing the research gaps improves our understanding of variability in Dutch children's vocabulary outcomes, and the role of dyadic and multimodal behaviours of caregiver-infant interactions in explaining part of this variation. We introduce each study and the research questions below.

1.5. Chapter overviews and research questions

To address the methodological and theoretical gaps in the literature, we analyse the performance of open-source automated tools on an existing corpus of Dutch IDS in

Chapter 2. We analyse concurrent and longitudinal vocabulary outcomes obtained via caregiver reports, when the children were around 10 months and around 3 years of age, and a lab-administered task of children's receptive vocabularies when the children were around 3 years of age in Chapters 3 and 5. We analyse video data from caregiver-infant interactions when the infants were around 10 months of age in Chapters 4 and 5. Working with longitudinal data from a large cohort study provides us with unique opportunities to study a large sample of caregiver-infant dyads using multiple vocabulary outcomes collected across children's development.

Chapter 2 explores a methodological issue related to data annotation. We aim to examine the effectiveness of open-source ASR tools for the annotation of Dutch IDS. Research with infants is often limited to small samples because it is difficult, expensive, and time-consuming to recruit infants (see Oakes, 2017). Large-scale cohort studies provide us with much larger research samples compared to what is normally feasible within single research projects. However, such large samples come with another challenge: We still need to manually annotate all the raw data. In many languages, open-source ASR tools have been trained on ADS which could facilitate the manual annotation process. However, the speech register IDS has specific acoustic properties compared to ADS, such as a higher mean pitch, a larger pitch range, and a slower speech rate, which might pose a challenge for ASR tools developed for ADS. While these acoustic properties facilitate children's word recognition abilities (e.g., Estes & Hurley, 2013; Singh et al., 2009; Song et al., 2010; Zangl et al., 2005), they might hinder the performance of ASR tools which acoustic models are trained on ADS (Kirchhoff & Schimmel, 2005). To examine the accuracy of open-source ASR tools for the annotation of Dutch IDS, we need an annotated corpus of Dutch IDS. The caregiver-infant interactions collected in the YOUth cohort study are not yet transcribed at the level necessary for ASR. Therefore, we will use the Dutch part of the cross-linguistic corpus of Dutch and Mandarin Chinese IDS for this chapter (Han, 2019). Comparing the automatic annotations to the manual annotations allows us to assess the accuracy of the ASR tools for the annotation of Dutch IDS. If we can successfully use open-source ASR tools to facilitate the annotation process of IDS,

this could drastically speed up research in this area. Thus, the first research question addressed in Chapter 2 is:

> **RQ 1: To what extent can we use open-source ASR tools to successfully transcribe Dutch IDS?**

In Chapter 3, we first aim to address the second research gap related to caregiver reporting of Dutch children's vocabularies. We assess the validity and reliability of the N_{YOUth}-CDIs before we continue to use them to examine individual variability across children's vocabularies. We achieve this by calculating the reliability of each scale included in the N_{YOUth}-CDIs separately, as well as determining the concurrent and predictive validity of each scale – also with another standardised task of children's receptive vocabulary (i.e., PPVT-III-NL). The second research question addressing a methodological issue is:

> **RQ 2: Are the N_{YOUth}-CDIs valid and reliable measures of Dutch children's vocabulary?**

Then, we also address the third gap concerning the onset and stability of well-known predictors of variation in children's vocabulary outcomes. We examine whether certain demographic effects on children's vocabulary, such as children's gender and maternal education, are age-specific and task-specific. While these factors have been studied extensively, they have rarely been studied within one large, longitudinal sample using different vocabulary outcome measures across children's development. This provides us with more insights into the stability of these effects over time while keeping the sample constant. The third research question is:

> **RQ 3: Are demographic predictors of variation in children's vocabularies age-specific and task-specific?**

Chapter 4 explores methodological issues and knowledge gaps concerning dyadic and multimodal behaviours during caregiver-infant interactions. Before we can assess the influences of dyadic and multimodal behaviours on children's vocabulary outcomes, we first have to develop and test a new coding scheme including infants' vocalisations and gestures and caregivers' verbal and nonverbal responses that might help infants disambiguate word-referent relations in the learning environment, such as gestures or body orientation. In this chapter, we present a characterisation of 9- to 11-month-old infants' vocalisations and gestures and their caregivers' verbal, nonverbal, and multimodal (i.e., verbal and nonverbal) responses during free play. Then, we examine whether different infants' vocalisations and gestures elicited different rates and types of verbal, nonverbal, and multimodal responses from caregivers. We examine this through statistical analyses of co-occurring infant behaviours and caregiver responses. This informs us to what extent infants influence their caregivers' responsiveness during caregiver-infant interactions, thereby shaping their own early learning environments. The research questions are:

RQ 4: What types of caregivers' verbal, nonverbal, and multimodal responses to infants' vocalisations and gestures do we observe during free play?

RQ 5: Do caregivers' verbal, nonverbal, and multimodal responses differ as a function of infants' vocalisations or gestures?

In Chapter 5, we assess whether different combinations of dyadic behaviours (i.e., coupled infant behaviours and caregiver verbal and multimodal responses) are better predictors of children's concurrent and longitudinal vocabulary outcomes than infants' individual behaviours. Previous studies have largely focused on the contributions of infants' individual behaviours, particularly infant gestures, in explaining variation in children's vocabulary development (e.g., Brooks & Meltzoff, 2008; Colonnesi et al., 2010; Rowe & Goldin-Meadow, 2009). Other studies have

shown that caregivers' responses to the infants' behaviours are correlated with child vocabulary (e.g., Bornstein et al., 2008; McGillion et al., 2013; Olson & Masur, 2015; Tamis-LeMonda et al., 2001; Wu & Gros-Louis, 2015), and they appear to mediate the relationship between infants' gestures and child vocabulary (see Olson & Masur, 2015). If infants' behaviours predict children's vocabulary outcomes because they tend to elicit contingent responses from caregivers during interactions, we expect that the dyadic combinations of infants' behaviours and caregivers' responses explain more variation in children's vocabularies. We also expect that caregivers' multimodal responses facilitate children's vocabulary development. If so, this would highlight the importance of studying dyadic and multimodal combinations of behaviours during real-time caregiver-infant interactions and adds to our understanding of the mechanisms underlying the facilitative role of infants' vocalisations and gestures in their vocabulary development. The last research question is:

> **RQ 6: Do dyadic combinations of infants' vocalisations and gestures (shows+gives and points) and caregivers' verbal and multimodal responses during free play improve the predictive value of infants' behaviours for children's vocabulary outcomes?**

Chapter 6 contains a general discussion of all chapters. The research findings presented and discussed in this dissertation lead to recommendations for other researchers, highlight potential issues when studying large datasets, inform theory, and provide us with new directions for future studies.

1.6. Publication status of chapters

- Chapter 2: van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2023). Using open-source automatic speech recognition tools for the annotation of Dutch infant-directed speech. *Multimodal Technologies and Interaction*, *7*(7), 68. https://doi.org/10.3390/mti7070068

- Chapter 3: van der Klis, A., Junge, C., Adriaans, F., & Kager, R. (submitted). Caregiver reports of Dutch children's vocabularies: Effects on vocabulary are age-specific and task-specific.
- Chapter 4: van der Klis, A., Adriaans, F., & Kager, R. (2023). Infants' behaviours elicit different verbal, nonverbal, and multimodal responses from caregivers during early play. *Infant Behavior and Development*, *71*, 101828.
- Chapter 5: van der Klis, A., Junge, C., Adriaans, F., & Kager, R. (submitted). The role of dyadic coordination of infants' behaviours and caregivers' verbal and multimodal responses in predicting vocabulary outcomes.

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*(2), 484–501. https://doi.org/10.1111/cdev.13508

Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development*, *60*(2), 381–398. https://doi.org/10.1111/j.1467-8624.1989.tb02723.x

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, *67*(6), 3135–3153. https://doi.org/10.2307/1131771

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5–22. https://doi.org/10.1016/S0167-6393(00)00067-4

Bavin, E. L., Prior, M., Reilly, S., Bretherton, L., Williams, J., Eadie, P., Barrett, Y., & Ukoumunne, O. C. (2008). The Early Language in Victoria Study: Predicting vocabulary at age one and two years from gesture and object use. *Journal of Child Language*, *35*(3), 687–701. https://doi.org/10.1017/S0305000908008726

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Blake, J., O'Rourke, P., & Borzellino, G. (1994). Form and function in the development of pointing and reaching gestures. *Infant Behavior & Development*, *17*(2), 195–203. https://doi.org/10.1016/0163-6383(94)90055-8

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476. https://doi.org/10.1017/S0142716416000060

Bornstein, M. H., & Haynes, O. M. (1998). Vocabulary competence in early childhood: Measurement, latent construct, and predictive validity. *Child Development*, *69*(3), 654–671. https://doi.org/10.1111/j.1467-8624.1998.tb06235.x

Bornstein, M. H., & Putnick, D. L. (2012). Stability of language in childhood: A multiage, multidomain, multimeasure, and multisource study. *Developmental Psychology*, *48*(2), 477–491. https://doi.org/10.1037/a0025889

Bornstein, M. H., Tamis-LeMonda, C. S., Hahn, C.-S., & Haynes, O. M. (2008). Maternal responsiveness to young children at three ages: Longitudinal analysis of a multidimensional, modular, and specific parenting construct. *Developmental Psychology*, *44*(3), 867–874. https://doi.org/10.1037/0012-1649.44.3.867

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 and Development*, *22*(1), 65–85. https://doi.org/10.1016/S0163-6383(99)80006-X

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2019). Intention or attention before pointing: Do infants' early holdout gestures reflect evidence of a declarative motive? *Infancy*, *24*(2), 228–248. https://doi.org/10.1111/infa.12267

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': Modifications in mothers' infant-directed action. *Developmental Science*, *5*(1), 72–83. https://doi.org/10.1111/1467-7687.00211

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/s030500090700829x

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

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), i–vi, 1–143.

Chen, C., Houston, D. M., & Yu, C. (2021). Parent–child joint behaviors in novel object play create high-quality data for word learning. *Child Development*, *92*(5), 1889–1905. https://doi.org/10.1111/cdev.13620

Colonnesi, C., Stams, G. J. J. M., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366. https://doi.org/10.1016/j.dr.2010.10.001

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345. https://doi.org/10.1111/cdev.12193

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2019). Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants' vocalizations, gestures and word production. *Developmental Science*, *23*(1), 1–21. https://doi.org/10.1111/desc.12843

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-III*. American Guidance Service.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1), 1–13.

Estes, K. G., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Fabes, R. A., Gaertner, B. M., & Popp, T. K. (2006). Getting along with others: Social competence in early childhood. In K. McCartney & D. Phillips, *Blackwell handbook of early childhood development* (pp. 296–316). Blackwell Publishing. https://doi.org/10.1002/9780470757703.ch15

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*(2), 310–322. https://doi.org/10.1111/1467-8624.00146

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. https://doi.org/10.2307/1166093

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur Communicative Development Inventories: User's guide and technical manual* (Second edition). Paul H. Brookes Publishing Co.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior & Development*, *8*(2), 181–195. https://doi.org/10.1016/S0163-6383(85)80005-9

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank Project*. MIT Press. https://langcog.github.io/wordbank-book/

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621–632. https://doi.org/10.1037/0012-1649.40.4.621

Ger, E., Altınok, N., Liszkowski, U., & Küntay, A. C. (2018). Development of infant pointing from 10 to 12 months: The role of relevant caregiver responsiveness. *Infancy*, *23*(5), 708–729. https://doi.org/10.1111/infa.12239

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*(4), 878–894. https://doi.org/10.1111/1467-8624.00197

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Goldstein, M. H., Schwade, J. A., & Bornstein, M. H. (2009). The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child Development*, *80*(3), 636–644. https://doi.org/10.1111/j.1467-8624.2009.01287.x

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science*, *13*(1), 252–263. https://doi.org/10.1111/j.1467-7687.2009.00871.x

Han, M. (2019). *The role of prosodic input in word learning: A cross-linguistic investigation of Dutch and Mandarin Chinese infant-directed speech* [Dissertation, Utrecht University]. http://localhost/handle/1874/379614

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children* (pp. xxiii, 268). Paul H Brookes Publishing.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

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. https://doi.org/10.1016/j.cogpsych.2010.08.002

Hymes, D. (1972). On communicative compentence. In J. Pride & J. Holmes, *Sociolinguistics* (pp. 269–285). Penguin Books.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., & Jasnow, M. D. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development*, *66*(2), i–viii, 1–132.

Junge, C., Valkenburg, P. M., Deković, M., & Branje, S. (2020). The building blocks of social competence: Contributions of the Consortium of Individual Development. *Developmental Cognitive Neuroscience*, *45*, 100861. https://doi.org/10.1016/j.dcn.2020.100861

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*(9), 323–328. https://doi.org/10.1016/S1364-6613(99)01363-7

Jusczyk, P. W., Hirsh-Pasek, K., Nelson, D. G., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, *24*(2), 252–293. https://doi.org/10.1016/0010-0285(92)90009-q

Keller, H., Lohaus, A., Völker, S., Cappenberg, M., & Chasiotis, A. (1999). Temporal contingency as an independent component of parenting behavior. *Child Development*, *70*(2), 474–485.

Kidd, E., & Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics*, *6*(1), 319–340. https://doi.org/10.1146/annurev-linguistics-011619-030326

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4 Pt 1), 2238–2246. https://doi.org/10.1121/1.1869172

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Lee, J. (2011). Size matters: Early vocabulary as a predictor of language and literacy competence. *Applied Psycholinguistics*, *32*(1), 69–92. https://doi.org/10.1017/S0142716410000299

Liszkowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science*, *7*(3), 297–307. https://doi.org/10.1111/j.1467-7687.2004.00349.x

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In J. P. Byrnes & S. A. Gelman (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 72–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511983689.004

Masek, L. R., McMillan, B. T. M., Paterson, S. J., Tamis LeMonda, C. S., Golinkoff, R. M., & Hirsh-Pasek, K. (2021a). Where language meets attention: How contingent interactions promote learning. *Developmental Review*, *60*, 100961. https://doi.org/10.1016/j.dr.2021.100961

Masek, L. R., Ramirez, A. G., McMillan, B. T. M., Hirsh-Pasek, K., & Golinkoff, R. M. (2021b). Beyond counting words: A paradigm shift for the study of language acquisition. *Child Development Perspectives*, *15*(4), 274–280. https://doi.org/10.1111/cdep.12425

May, L., Byers-Heinlein, K., Gervain, J., & Werker, J. F. (2011). Language and the newborn brain: Does prenatal language experience shape the neonate neural response to speech? *Frontiers in Psychology*, *2*, 222. https://doi.org/10.3389/fpsyg.2011.00222

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

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631. https://doi.org/10.1126/science.1144073

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178. https://doi.org/10.1016/0010-0277(88)90035-2

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: A two-country study. *Acta Paediatrica*, *102*(2), 156–160. https://doi.org/10.1111/apa.12098

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). 24-month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child Development*, *86*(5), 1351–1370. https://doi.org/10.1111/cdev.12398

Naigles, L. R., & Swensen, L. D. (2007). Syntactic supports for word learning. In E. Hoff & M. Shatz, *Blackwell handbook of language development* (pp. 212–231). Blackwell Publishing. https://doi.org/10.1002/9780470757833.ch11

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science*, *6*(2), 136–142. https://doi.org/10.1111/1467-7687.00263

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186

Olson, J., & Masur, E. F. (2011). Infants' gestures influence mothers' provision of object, action and internal state labels. *Journal of Child Language*, *38*(5), 1028–1054. https://doi.org/10.1017/S0305000910000565

Olson, J., & Masur, E. F. (2013). Mothers respond differently to infants' gestural versus nongestural communicative bids. *First Language*, *33*(4), 372–387. https://doi.org/10.1177/0142723713493346

Olson, J., & Masur, E. F. (2015). Mothers' labeling responses to infants' gestures predict vocabulary outcomes. *Journal of Child Language*, *42*(6), 1289–1311. https://doi.org/10.1017/S0305000914000828

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., van der Wal, J. J. M., Hulshoff Pol, H. E., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, *46*, 100868. https://doi.org/10.1016/j.dcn.2020.100868

Pan, B. A., Rowe, M. L., Spier, E., & Tamis-Lemonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language*, *31*(3), 587–608. https://doi.org/10.1017/S0305000904006270

Preston, J. L., Frost, S. J., Mencl, W. E., Fulbright, R. K., Landi, N., Grigorenko, E., Jacobsen, L., & Pugh, K. R. (2010). Early and late talkers: School-age language, literacy and neurolinguistic differences. *Brain*, *133*(8), 2185–2195. https://doi.org/10.1093/brain/awq163

Quine, W. V. (1964). *Word and object*. MIT Press.

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur Communicative Development Inventory: Words and Sentences. *Journal of Child Language*, *27*(2), 255–266. https://doi.org/10.1017/S0305000900004098

Renzi, D. T., Romberg, A. R., Bolger, D. J., & Newman, R. S. (2017). Two minds are better than one: Cooperative communication as a new framework for understanding infant language learning. *Translational Issues in Psychological Science*, *3*, 19–33. https://doi.org/10.1037/tps0000088

Rose-Krasnor, L. (1997). The nature of social competence: A theoretical review. *Social Development*, *6*(1), 111–135. https://doi.org/10.1111/j.1467-9507.1997.tb00097.x

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., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rowe, M. L., & Weisleder, A. (2020). Language development in context. *Annual Review of Developmental Psychology*, *2*(1), 201–223. https://doi.org/10.1146/annurev-devpsych-042220-121816

Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL*. Hartcourt Assessment BV.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*, 654–666. https://doi.org/10.1080/15250000903263973

Skinner, E. A. (1986). The origins of young children's perceived control: Mother contingent and sensitive behavior. *International Journal of Behavioral Development*, *9*(3), 359–382. https://doi.org/10.1177/016502548600900307

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Soderstrom, M. (2019). ManyBabies1: Infants' preference for infant-directed speech. *The Journal of the Acoustical Society of America*, *145*(3), 1728–1728. https://doi.org/10.1121/1.5101348

Soderstrom, M., Blossom, M., Foygel, R., & Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, *35*(4), 869–902. https://doi.org/10.1017/S0305000908008763

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

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development*, *90*(4), 1303–1318. https://doi.org/10.1111/cdev.13072

Spinelli, M., Fasolo, M., & Mesman, J. (2017). Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Developmental Review*, *44*, 1–18. https://doi.org/10.1016/j.dr.2016.12.001

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*(2), 121–126. https://doi.org/10.1177/0963721414522813

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*, 929–938. https://doi.org/10.1037/0012-1649.44.4.929

Tomasello, M., & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in Child Development*, *59*, 174–179. https://doi.org/10.1111/j.1540-5834.1994.tb00186.x

Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C., Milan Maillo, I., & Perniss, P. (2019). Onomatopoeia, gestures, actions and words: How do caregivers use multimodal cues to communicate with their children. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, *41*, 1171–1177. https://doi.org/10/1/Toy_Task_CogSci_final.pdf

Vygotsky, L. (1962). *Thought and language*. MIT Press. https://doi.org/10.1037/11193-000

Wu, Z., & Gros-Louis, J. (2015). Caregivers provide more labeling responses to infants' pointing than to infants' object-directed vocalizations. *Journal of Child Language*, *42*(3), 538–561. https://doi.org/10.1017/S0305000914000221

Yurovsky, D. (2018). A communicative approach to early word learning. *New Ideas in Psychology*, *50*, 73–79. https://doi.org/10.1016/j.newideapsych.2017.09.001

Zangl, R., Klarman, L., Thal, D., Fernald, A., & Bates, E. (2005). Dynamics of word comprehension in infancy: Developments in timing, accuracy, and resistance to acoustic degradation. *Journal of Cognition and Development*, *6*(2), 179–208. https://doi.org/10.1207/s15327647jcd0602_2

Zink, I., & Lejaegere, M. (2002). *N-CDI's: Lijsten voor Communicatieve Ontwikkeling. Aanpassing en hernormering van de MacArthur CDI's van Fenson et al.* Acco.

Zink, I., & Lejaegere, M. (2003). *N-CDI's: Korte vormen, Aanpassing en hernormering van de MacArthur Short Form Vocabulary Checklist van Fenson et al.* Acco.

Chapter 2

Using open-source automatic speech recognition tools for the annotation of Dutch infant-directed speech

Abstract

There is a large interest in the annotation of speech addressed to infants. Infant-directed speech (IDS) has acoustic properties that might pose a challenge to automatic speech recognition (ASR) tools which have been developed for adult-directed speech (ADS). While ASR tools could potentially speed up the annotation process, their effectiveness on this speech register is currently unknown. In this study, we assessed to what extent open-source ASR tools can successfully transcribe IDS. We used speech data from 21 Dutch mothers reading picture books containing target words to their 18- and 24-month-old children (IDS) and the experimenter (ADS). In Experiment 1, we examined how the ASR tool Kaldi-NL performs at annotating target words in IDS vs. ADS. We found that Kaldi-NL only found 55.8% of target words in IDS, while it annotated 66.8% correctly in ADS. In Experiment 2, we aimed to assess the difficulties in annotating IDS more broadly by transcribing all IDS utterances manually and comparing the word error rates (WERs) of two different ASR systems: Kaldi-NL and WhisperX. We found that WhisperX performs significantly better than Kaldi-NL. While there is much room for improvement, the results show that automatic transcriptions provide a promising starting point for researchers who have to transcribe large amounts of speech directed at infants.

This chapter has been published as:

van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2023). Using open-source automatic speech recognition tools for the annotation of Dutch infant-directed speech. *Multimodal Technologies and Interaction*, *7*(7), 68.

2.1. Introduction

When addressing infants, adults spontaneously adopt a different speech register referred to as infant-directed speech (IDS) or baby talk (e.g., Fernald & Simon, 1984; Fernald et al., 1989; Kuhl et al., 1997). This speech register is characterised by a variety of intonational and prosodic characteristics, including a higher mean pitch, a larger pitch range, and greater pitch variability compared to adult-directed speech (ADS) (for a review, see Soderstrom, 2007). IDS has also been found to have a slower speech rate than ADS in many languages, including Dutch (Johnson et al., 2013; Han et al., 2021). Many studies have reported positive links between the acoustic properties of IDS and children's linguistic outcomes (for a meta-analysis, see Spinelli et al., 2017). The mechanisms driving this relationship are still widely debated.

Previous studies have shown that slow speech improves children's word recognition performance (Song et al., 2010; Zangl et al., 2005). Han et al. (2021) showed that Dutch mothers slowed down speech when introducing unfamiliar words compared to familiar words. The results are less conclusive for pitch. In Singh et al. (2009), 7- and 8-month-old infants were able to recognise words that were previously presented in IDS but not when they were presented in ADS. Similarly, Estes and Hurley (2013) showed that 17.5-month-old children only learned novel words in IDS but not in ADS. The effects of pitch have not been studied in isolation, thus it remains unclear whether the facilitative effects of pitch on word recognition can be attributed to pitch alone. In addition, Han et al. (2020) found that Dutch mothers increase pitch for familiar words, while Chinese mothers increase pitch for unfamiliar words. Pitch may function differently in these languages, and it remains unclear how pitch facilitates learning. It has also been suggested that infants prefer listening to IDS over ADS (Fernald, 1985; Dunst et al., 2012), indirectly facilitating the learning process.

The exaggerated prosody of IDS may also facilitate the learning of vowel categories. Kuhl et al. (1997) suggested that vowels in IDS are acoustically more extreme, containing larger vowel spaces than vowels produced in ADS, leading to this

hypothesis. Recent studies have reported that IDS vowels are produced with higher variability compared to vowels in ADS (Cristia & Seidl, 2014; Miyazawa et al., 2017), resulting in more overlap between vowel categories in IDS. Adriaans and Swingley (2017) proposed that among this high variability, however, mothers produce exaggerated high-quality instances of vowels which could facilitate vowel categorisation. The trade-off between larger vowel spaces, higher variability, smaller contrasts, and the presence of high-quality tokens in the input remains to be seen.

There seems to be a general trend of IDS becoming prosodically more like ADS as children grow older (e.g., Kitamura et al., 2001). Specifically, Han et al. (2020) found that utterance mean pitch was significantly lower when Dutch mothers addressed their 24-month-old infants compared to addressing their 18-month-old infants, even though utterance mean pitch was still higher in IDS than ADS at both ages. Sjons et al. (2017) found an increase in articulation rate of Swedish IDS from 7 to 33 months suggesting that the articulation rate of IDS becomes more similar to the articulation rate of ADS as children grow older. Nevertheless, IDS was still slower than ADS. Han et al. (2021), on the other hand, did not find any evidence for age-related changes in articulation rate in Dutch IDS from 18 to 24 months, and speaking rate remained slower compared to ADS. Age-related effects vary cross-linguistically, but there is a trend of IDS becoming acoustically more like ADS over time.

To identify and analyse the distinctive properties of IDS, and subsequently advance our understanding of the role of IDS in language development, it is essential to collect and transcribe IDS in many languages and across many speakers, for infants at different ages. Preparing speech data for analysis takes a notoriously long time. Segmenting, annotating, and transcribing an hour of speech, including verifying the quality of the transcription, can take up to fifty hours in total depending on the contents (Barras et al., 2001). Using automatic transcriptions may serve as a good starting point to diminish the time needed for manual annotation. Depending on the research aims and the accuracy needed to accomplish these, automatic transcriptions may be used as a starting point and then manually corrected by a human annotator to save time (see

Gaur et al., 2016). Tools are being developed to generate automatic annotations which would benefit research on IDS by speeding up the annotation process (e.g., Burnham et al., 2016). Currently, it is still common practice in the field to transcribe speech manually. To date, studies have not yet addressed to what extent we could use off-the-shelf ASR tools to facilitate the annotation process of IDS. The current study assessed the performance of ASR tools in the annotation of Dutch speech directed at 18-month-old and 24-month-old infants. In Experiment 1, we examined how the ASR tool Kaldi-NL performs at annotating target words in IDS vs. ADS. In Experiment 2, we examined the performance on IDS more broadly by testing two different ASR systems (Kaldi-NL and WhisperX) on the complete set of IDS utterances. We compared their performance in terms of word error rates (WERs). The experiments inform us to what extent off-the-shelf ASR tools trained on ADS are successful at annotating Dutch IDS.

2.1.1. Previous work

Automatic speech recognition (ASR) is the process of generating text representations for acoustic speech input. ASR systems have components which require extensive training, such as an acoustic model and a language model. The acoustic model learns from audio recordings combined with phonetic transcriptions, creating statistical representations of speech sounds. Many ASR systems are using deep neural networks to create these representations, drastically improving their automatic transcription performance (e.g., Mohamed et al., 2012). The acoustic model translates the audio signal into a sequence of the most probable phonemes. The language model learns from a large corpus of transcribed speech, creating statistical probabilities of word sequences in the language. The ASR system combines the two models to produce the most likely written transcription of the signal as output.

Previous studies have examined whether certain acoustic features can predict ASR errors. Fast speech and extremely long word durations are both related to higher error rates (e.g., Goldwater et al., 2010; Kawahara et al., 2003; Shinozaki et al., 2001). Goldwater et al. (2010) found that extreme values of pitch and intensity also increase

error rates. In addition, an analysis of two human-computer dialogue systems shows that misrecognised utterances are associated with pitch excursions, loudness, and longer duration (Hirschberg et al., 2004). The authors marked these as instances of hyperarticulated speech. Importantly, some of these features (i.e., above-average mean pitch and pitch range, below-average speech rate) are similar to the typical features of IDS.

Precisely because IDS is highly variable and exaggerated speech, it may also serve as good training data resulting in more robust models (Kirchhoff & Schimmel, 2005; Shinozaki et al., 2009). The acoustic characteristics of ID – potentially resulting in phonetic categories that are well separated in the input space – could also aid phonetic classification using Gaussian mixture models. Kirchhoff and Schimmel (2005) trained an ASR system on IDS, using recordings of 22 American English mothers addressing their 2- to 5- month-old infants, and another one on ADS, using the same mothers addressing the adult experimenter. The ADS-trained system was highly accurate at recognising target words in ADS (95.5%) but less in IDS (81.6%). The system trained on IDS was notably better at recognising target words in IDS (93.5%), but the IDS-trained system did not perform as well on recognising the same target words in ADS (90.2%) (Kirchhoff & Schimmel, 2005). These results indicate that a matched system (i.e., trained and tested on the same speech register) produces the best recognition results. The largest degradation in performance is found when an ASR system trained on ADS is used for the recognition of IDS which is the more variable speech register. Nevertheless, the authors used a relatively small set of training data (utterances by 22 speakers). Currently, we do not know whether an ASR system trained on a much larger ADS data set contains more robust models that are more suitable for the recognition of IDS.

2.2. Experiment 1

In the first experiment, we aimed to assess to what extent an open-source ASR tool, Kaldi-NL, was successful at annotating target words in continuous, semi-naturalistic IDS. This is the first study to 1) address this question for Dutch, 2) use a readily

available open-source ASR tool, 3) compare the recognition performance of IDS addressed to different age groups, and 4) examine the effects of different acoustic features (i.e., mean pitch, pitch range, and articulation rate) on recognition accuracy. While acoustic deviations in pitch and speech rate support children's word recognition abilities (e.g., Estes & Hurley, 2013; Singh et al., 2009; Song et al., 2010; Zangl et al., 2005), previous studies have shown that these may have negative effects on ASR performance (e.g., Goldwater et al., 2010; Kawahara et al., 2003; Shinozaki et al., 2001). Given that ASR performance is negatively affected by acoustic deviations, and that Dutch IDS is marked by a higher mean pitch and slower speech rate compared to Dutch ADS, we would expect that an ASR system trained on Dutch ADS exhibits lower performance when transcribing Dutch IDS. Very few studies have assessed the accuracy of ASR systems at transcribing IDS, and none so far for Dutch. It is important to verify whether research findings generalise to other languages. First, we compared the accuracy of Kaldi-NL at transcribing target words produced by Dutch mothers in continuous IDS directed at 18-month-old children and 24-month-old children and the same target words produced in ADS. Then, we examined which acoustic features affected speech recognition accuracy using a logistic mixed-effects model. The results informed us to what extent an off-the-shelf ASR tool can successfully transcribe IDS and whether the transcription accuracy is negatively affected by IDS.

2.3. Materials and methods

2.3.1. Participants

This study is part of a larger cross-linguistic corpus of Dutch and Mandarin Chinese infant-directed speech (Han, 2019). The speech data collection methods are identical to those reported in Han et al. (2020, 2021). From this corpus, we included 21 Dutch-speaking mother-child dyads who were recruited from the Utrecht Baby Lab database and were all Dutch native speakers living in the Utrecht area in the Netherlands. We used a longitudinal design and collected mothers' ADS and IDS speech data when their children were 18 months old (M = 1;6, range = 1;6 – 1;7) and 24 months old (M = 2;1, range = 2;0 – 2;3). All mothers were native speakers of Dutch who followed

higher education (undergraduate degree and above). All children were typically developing with no report of language or hearing problems. All participating mothers signed informed consent forms.

2.3.2. Materials and procedure

Mothers read the same picture book to their infant, to elicit IDS, and to the adult experimenter (female), to elicit ADS, during the recording sessions. Different picture books for each time point (children's ages 18 months and 24 months) were designed to elicit two different sets of seven disyllabic target words. On each page, the target word was on the left side and an illustration including a depiction of the word was on the right side (for the picture book, see Han, 2019, p. 187). The mothers were instructed to tell the story including the target words, eliciting semi-naturalistic speech. The target words at both time points can be found in Table 2.1. These target words were selected because they were likely unfamiliar to the child (apart from 'apple' and 'grandpa' which were used for comparison) which was relevant to the previous study. The unfamiliar target words at 24 months are of much lower frequency than the unfamiliar target words at 18 months.

In total, the participants produced 1051 target words embedded in semi-naturalistic speech across both speech registers and time points. All mothers produced each target word at least once in each condition at each age. The productions are equally distributed: 563 target word productions when the infants were 18 months old (243 in ADS; 320 in IDS) and 488 target word productions when the infants were 24 months old (215 in ADS; 273 in IDS). The total duration of the speech sample was 97.95 minutes (ADS: 36.48 min; IDS: 61.47 min) at 18 months and 102.35 minutes (ADS: 35.65 min; IDS: 66.70 min) at 24 months. All participants were tested in a quiet room in the Utrecht Baby Lab. The audio recordings were made using a ZOOM H1 recorder with 16-bit resolution and a sampling rate of 44.1 kHz.

Table 2.1. Target words and their word frequencies according to the SUBTLEX corpus of Dutch (Keuleers et al., 2010).

18 months			24 months		
Dutch	Translation	Frequency	Dutch	Translation	Frequency
opa	"grandpa"	2507	opa	"grandpa"	2507
appel	"apple"	446	appel	"apple"	446
eland	"moose"	115	emoe	"emu"	6
bever	"beaver"	128	wezel	"weasel"	90
walnoot	"walnut"	31	bamboe	"bamboo"	30
kasteel	"castle"	1207	kapel	"chapel"	194
pompoen	"pumpkin"	109	jasmijn	"jasmine"	37

2.3.3. Transcriptions

We compared the automatic annotations to the manual annotations of target words to assess the accuracy of the ASR system at annotating target words in IDS. All target words were manually annotated in previous work (for details, see Han, 2019). A trained Dutch native speaker extracted the target words from the audio recordings using Praat (Boersma & Weenink, 2020). For the current study, the full recordings were automatically transcribed using the online Kaldi-NL ASR tool developed by the Dutch Foundation of Open Speech Technology and hosted by the Radboud University (Version 0.5.0; Yilmaz & van Gompel, 2020). The Dutch models have been developed at the University of Twente using the Spoken Dutch Corpus ("*Corpus Gesproken Nederlands*") containing about 900 hours of Dutch speech recordings from, for example, conversations and television shows (Oostdijk, 2000). Kaldi-NL has a lexicon of ca. 250 thousand words and employs time-delay neural network (TDNN) layers which have been shown to outperform low frame rate bidirectional long short-term memory acoustic models (see Peddinti et al., 2018). A recent study used Kaldi-NL to transcribe Dutch doctor-patient consultation recordings and found a WER of 25.8% without fine-tuning the language model or lexicon to include domain-specific healthcare words (Tejedor-García et al., 2022). To generate the

automatic transcriptions, the online system takes audio files (e.g., WAV) as input. After a short period of processing, the output of the ASR system consists of a plain text file containing a written transcription and a CTM file containing all transcribed words and their corresponding timestamps (i.e., indicating when it occurred in the audio file). Using this output, we lastly examined the accuracy of the automatic annotations of target words using the evaluation procedure described below.

2.3.4. Evaluation procedure

We compared the automatic annotations of the target words from the time-stamped CTM file to the manual annotations (i.e., the ground truth) using an interactive Python script. For each target word in the manual annotations, the script shows the timestamp of the word from the Praat TextGrid and the timestamp of the target word from the automatic transcription, in addition to playing both extractions from the audio file. The evaluator checks whether the manual and automatic words match. If yes, this counts as one "hit" (correctly identified target word). If not, then the word is a "miss" (not identified target word). Lastly, the script collects all target words that were automatically transcribed but not matched to a manual annotation and marked these as "false positives" (words incorrectly identified as target words). These cases were double-checked since the target words could potentially have been overlooked during the manual annotation process. The output of the script is a data file containing all assessed target words, the speech register (IDS or ADS), the time point (18m or 24m), the assessment (hit, miss, or false positive), and the timestamps from the TextGrid and from Kaldi-NL. All morphological varieties of the target words, such as diminutives (e.g., *appeltje* or *walnootje*), were also included in the data. We only analysed target words, and not full sentences, because it is easier to compare the data across speech registers and eliminates the chance that any observed differences between IDS and ADS can be attributed to the language model (e.g., IDS tends to have shorter sentences and more repetitions) or the vocabulary size (e.g., IDS tends to have shorter, simplified words).

The frequencies of hits, misses, and false positives allow us to calculate three common accuracy scores: recall, precision, and *F*-scores. Recall informs us how many of the total target words annotated manually were also found by the ASR system. Precision informs us how many of the recalls were target words, and not false positives. *F*-score is the harmonic mean between recall and precision (Goutte & Gaussier, 2005). This is an important additional measure because high recall does not equal high accuracy when precision is low, and vice versa. The measures are calculated as follows:

$$recall = \frac{hits}{hits + misses}$$

$$precision = \frac{hits}{hits + false\ positives}$$

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

2.3.5. Acoustic features of target words

We examined each target word's mean pitch, pitch range, and articulation rate. First, we automatically extracted the minimum pitch, maximum pitch, and mean pitch from each target word in IDS and ADS using a pitch range of 100-600 Hz in Praat (Boersma & Weenink, 2020). The top and bottom 5% of pitch measurements were all manually checked for pitch jumps (i.e., halving or doubling). In the case of a pitch jump, the pitch range was slightly adjusted to better fit the data. The pitch range was calculated by subtracting the minimum pitch from the maximum pitch of a target word. The articulation rate was calculated by dividing the number of syllables by the total duration of the target word. Target words were excluded when pitch could not be measured due to interference of the child's voice (n = 17) or due to whispering (n = 2), resulting in a final set of 1032 target words for the acoustic analysis.

2.3.6. Statistical analysis

To assess what affects ASR performance, we examined the effects of speech register, infant age, and the different acoustic properties of IDS – mean pitch, pitch range, and articulation rate – on recognition accuracy. The results of 1032 target words were analysed by fitting logistic mixed-effects models using the *lme4* package version 1.1-30 (Bates et al., 2015) in *R* version 4.2.0 (R Core Team, 2022) to predict recognition accuracy for each target word (hit or miss). The continuous variables F_0 mean, F_0 range, and articulation rate were centred and scaled. We used dummy coding for the dichotomous variables speech register (IDS or ADS) and age (18m or 24m) with ADS and 18m as reference levels. We added random intercepts for participants to account for potential individual variation in speech perceptibility and items because the target words were not the same across both time points and differed in word frequency. This can negatively impact recognition performance in a way that is not related to the measures that are of interest in the present study. Lastly, we calculated odds ratios from the regression coefficients to examine the impact of the predictors.

2.4. Results

We first calculated recall, precision, and *F*-scores to assess the accuracy of the ASR system for each speech register at each time point. This allowed us to compare the recognition accuracy for IDS to ADS. Then, we examined the distributions of the various acoustic measures across all conditions to examine whether the acoustic features that are typical of IDS – mean pitch, pitch range, and articulation rate – affected recognition accuracy. Lastly, we fitted a logistic mixed-effects model to examine which of the predictors has a significant effect on recognition accuracy.

2.4.1. Accuracy scores

For speech addressed to 18-month-old infants, the ASR system correctly annotated 180 of 320 (56.3%) target words. For ADS, the ASR system found 174 of 243 (71.6%) target words. For the 24-month-old infants, the system correctly annotated 151 of 273 (55.3%) target words in IDS and 132 of 215 (61.4%) target words in ADS. The difference in recognition accuracy between ADS and IDS diminished between the two

time points. The recall scores are visualised in Figure 2.1. All target words were correctly annotated at least once, indicating that none of the target words were out-of-vocabulary (i.e., all target words are present in Kaldi-NL's vocabulary).

In both registers, precision is 100%. Precision is calculated using false positives, and there were none in the data. For false positives to occur, other produced words must be phonologically similar to target words, which is unlikely given the limited contents of the picture books used in the present study. Table 2.2 contains the results of the evaluation procedure.

Figure 2.1. The proportions of hits and misses for each speech register within each age group.

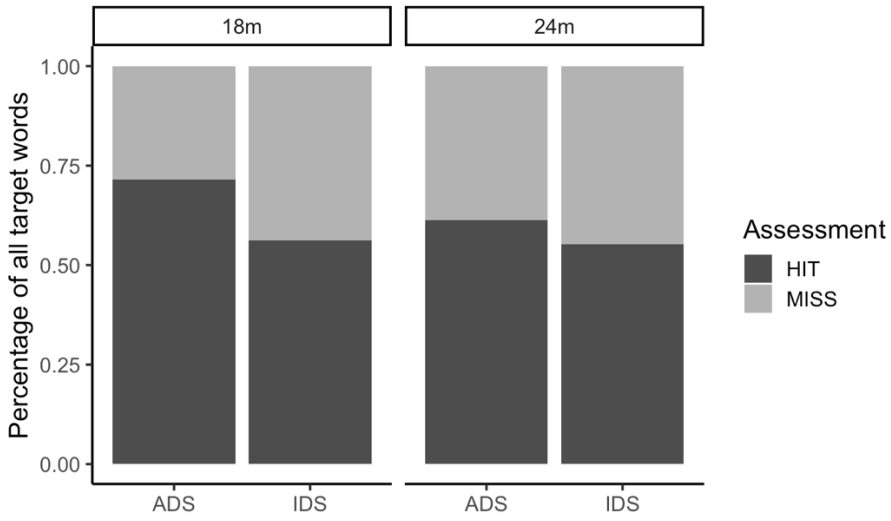


Table 2.2. Results of the evaluation procedure in proportions.

	18 months		24 months	
Register	ADS	IDS	ADS	IDS
Recall	0.72	0.56	0.61	0.55
Precision	1.00	1.00	1.00	1.00
F-score	0.84	0.72	0.76	0.71

Recall scores are generally lower for target words at 24 months, also for ADS. This is likely caused by the lower word frequencies of the target words that were produced at this age, as shown in Table 2.2. Low-frequency words have low probabilities in the language model of the ASR tool, making them less likely candidates to be selected. Therefore, the general word frequencies of the target words will affect recognition accuracy. The important finding is that the decrease in recognition accuracy found for IDS compared to ADS has become much smaller.

2.4.2. Acoustic measures

Figure 2.2 shows boxplots of the mean pitch of hits and misses in both speech registers. First, the boxplots show that on average, target words in IDS have a higher mean pitch than target words in ADS at both time points. Target words have the highest mean pitch in IDS at 18 months. Secondly, missed target words have on average a higher mean pitch than hits at 18 months. This difference seems to have disappeared at 24 months, although missed target words seem to have more extreme mean pitch values in both directions.

Figure 2.2. Boxplots depicting the lower (Q1) and upper (Q3) quartiles, the median, the minimum and maximum values, and outliers of the F*0* mean of target words.

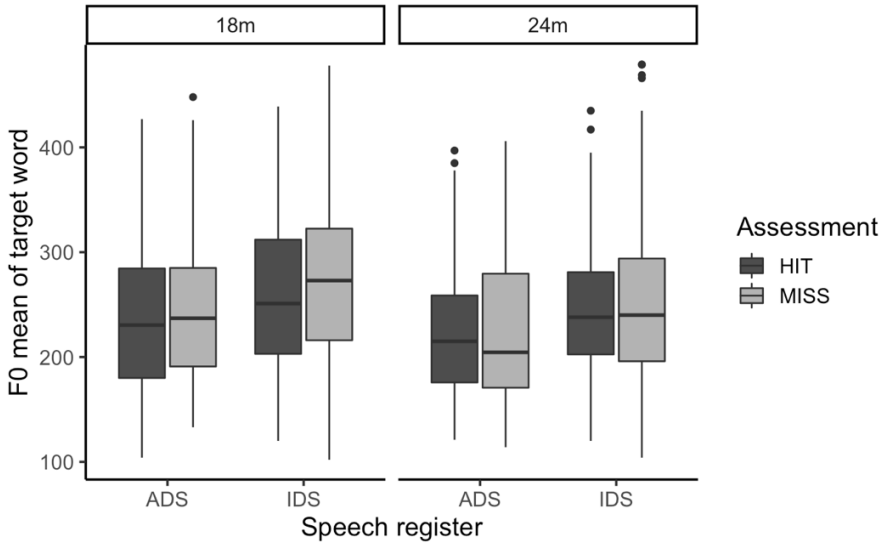


Figure 2.3 shows boxplots depicting pitch ranges of target words. First, the boxplots show that target words in IDS have a larger pitch range on average compared to target words in ADS. The figure does not provide evidence that missed target words have larger pitch ranges than hits on average.

Figure 2.4 shows boxplots depicting articulation rates. First, target words at 24 months are on average produced faster than target words at 18 months. At 24 months, target words in IDS were produced slower than target words in ADS. The difference between IDS and ADS is surprisingly smaller at 18 months, while we would expect IDS to become more similar to ADS over time. One explanation could be that the target words are of much lower frequency at 24 months, and mothers may lower their articulation rates more for didactic purposes when presenting unfamiliar words to their children (Han et al., 2021). Articulation rate does not seem to have a large effect on recognition accuracy, although we find more extreme values of low articulation rates across missed target words.

Figure 2.3. Boxplots depicting the lower (Q1) and upper (Q3) quartiles, the median, the minimum and maximum values, and outliers of the F0 range of target words.

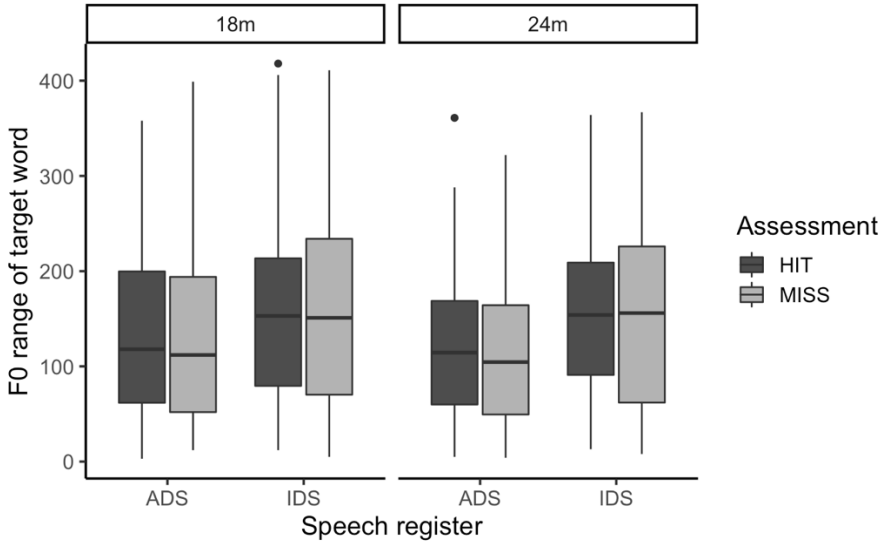
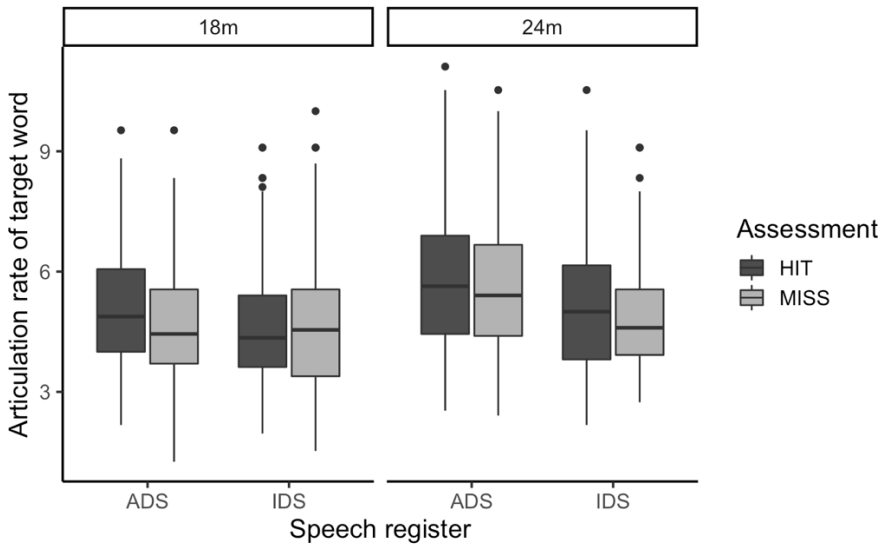


Figure 2.4. Boxplots depicting the lower (Q1) and upper (Q3) quartiles, the median, the minimum and maximum values, and outliers of the articulation rate (syllables/s) of target words.



The results of 1032 target words were analysed by fitting logistic mixed-effects models using a bottom-up approach. First, we found that adding random intercepts for participants and items significantly improved model fit. This indicates there is random variability across participants and items which affects recognition accuracy. Then, we examined which of the fixed effects (speech register, age, mean pitch, pitch range, and articulation rate) significantly improved model fit. We found that speech register and mean pitch significantly improved the fit of the model. There is a significant effect of age if we do not add a random intercept for item to the model. The effect of age likely disappears when adding a random intercept for item since the items differed across the two measurement waves, partially accounting for this effect. The model which includes a random intercept for item better fits the data. There was no improvement of model fit when adding an interaction between the fixed effects. The final model, including fixed effects for speech register and mean pitch and random intercepts for participant and item, is presented in Table 2.3.

Table 2.3. Results of the logistic mixed-effects model transformed to exponentiated coefficients (accuracy ~ speech register + F0 mean + (1|subject) + (1|item)).

Predictor	Exp. coefficient	*SE*	*Z*-value	*p*-value
(Intercept)	0.40	0.37	-2.51	0.01
Register (IDS)	1.86	0.15	4.08	0.00
F_0 mean	1.20	0.08	2.24	0.02

The speech register IDS is a strong predictor of a recognition error (i.e., a missed target word) ($p < 0.001$). When the target word is produced in IDS, there is an increase of 86% (95% CI [1.38, 2.51]) in the odds of the ASR system missing the target word compared to a target word produced in ADS. On top of this, there is a significant negative effect of mean pitch on recognition accuracy ($p = 0.02$). A one-unit increase in mean pitch results in an increase of 20% (95% CI [1.02, 1.40]) in the odds of the

ASR system missing a target word. Words produced with a higher mean pitch are problematic for the recognition of target words in continuous speech.

2.5. Discussion

The first aim of Experiment 1 was to assess the performance of an ASR tool at annotating target words in IDS vs. ADS. The results show that there is a large gap between the recognition accuracy of ADS and IDS, especially for speech directed at younger infants. Previous studies on IDS have shown that the acoustic features of IDS become less salient over time (e.g., Han et al., 2020; Kitamura et al., 2001). This could explain why the difference between IDS and ADS automatic recognition accuracy has become smaller for speech addressed to 24-month-olds compared to 18-month-olds. As we expected based on a previous study on American English, an ASR tool trained on ADS is less successful at transcribing IDS than ADS (Kirchhoff & Schimmel, 2005). The difficulties with transcribing IDS generalise to Dutch. While Kaldi-NL is trained on a significantly larger data set compared to the ASR system trained by Kirchhoff and Schimmel (2005) (i.e., 900 hours of speech compared to a set of utterances by 22 speakers), this did not help much to overcome the difficulties with recognising this speech register.

We also examined which factors are predictors of a recognition error made by Kaldi-NL. The results show that IDS as a speech register is an important predictor of a missed target word. We also found a significant negative effect of mean pitch on recognition accuracy. Previous studies found that slow speech facilitates word recognition in children (Song et al., 2010; Zangl et al., 2005), although it could hinder ASR performance (Goldwater et al., 2010; Kawahara et al., 2003; Shinozaki et al., 2001). We did not find a significant effect of articulation rate on ASR accuracy. It could be possible that the ASR system trained on ADS does not have difficulties with the larger pitch range or slower articulation rate of IDS, but the ASR system does show a decrease in accuracy when transcribing target words with a higher mean pitch. The results of the mixed-effects model suggest that IDS is also likely more difficult to be recognised by the ASR system for reasons beyond the examined acoustic

measures, for example, due to the high amount of acoustic variability or any syntactic differences.

2.5.1. Follow-up experiment

We found that Kaldi-NL transcribed approximately half of the target words correctly in IDS and approximately two-thirds in ADS. Based on this performance, we asked two follow-up questions. First, the target words, while important to the previous experiment, constitute only a relatively small portion of the total set of words in the data. One important question is thus to what extent do the results of target words generalise to the automatic transcription of full sentences? Second, since we tested only one particular system in Experiment 1, the question is to what extent the performance is reflective of ASR systems in general. That is, to what extent are the recognition results similar across different ASR systems? In Experiment 2, we tackle these two questions in parallel by manually transcribing all IDS utterances in the data set and comparing the two different systems (Kaldi-NL and the newly available open-source WhisperX) on their ability to transcribe these utterances, as measured by their word error rates (WERs). By analysing full sentences instead of target words, we have over twenty times more IDS data, while the data are less affected by the large differences in target word frequencies across the two time points.

2.6. Experiment 2

A previous study has found that when mothers are reading a picture book containing target words to their infants, mothers consistently positioned these words on exaggerated pitch peaks (Fernald & Mazzie, 1991). Mothers did not do this when reading the picture book to an adult. Therefore, the results of target words as opposed to full utterances could have inflated the negative effects of the speech register IDS on ASR performance. However, the findings by Fernald and Mazzie (1991) were not replicated in this Dutch IDS data set by Han et al. (2020). In Dutch IDS, the pitch of target words was similar to the pitch of utterances, although utterances are characterised by less variability than target words. Therefore, it is not likely that by examining target words alone, we have inflated the effects of the acoustics of the

speech register IDS on recognition accuracy. Nevertheless, the performance of ASR systems on full sentences in IDS remains to be seen.

In the previous experiment, we found that Kaldi-NL was less accurate at transcribing target words directed at younger children compared to older children. We hypothesised that because the acoustic features of IDS are more prominent when children are younger (e.g., Han et al., 2020; Kitamura et al., 2001), the ASR system trained on ADS has more difficulties transcribing speech addressed to younger children. The results for speech addressed to 18-month-old children and 24-month-old children in the previous experiment were difficult to compare since the target words examined at the two time points were of vastly different frequencies – influencing ASR performance. By calculating WERs of full utterances, we reduce the influence of target word frequencies on the results.

2.6.1. Research aim

In this second experiment, we first aimed to evaluate how open-source ASR systems perform at transcribing full utterances in Dutch IDS. We compared WERs of utterances in IDS directed at 18-month-old and 24-month-old infants. We expected that WERs are lower in speech directed at older infants since IDS becomes prosodically more similar to ADS as children grow older (e.g., Han et al., 2020; Kitamura et al., 2001). Based on the performance of Kaldi-NL in the previous experiment, the second aim was to assess whether an ASR system trained on a much larger, semi-supervised data set performs similarly at transcribing Dutch IDS. It might be possible that a larger training set results in more robust models that are more successful at transcribing a more variable speech register like IDS. We compared the WERs of two different ASR systems (Kaldi-NL and WhisperX) for the transcription of Dutch IDS. The second experiment informed us whether the results of target words in the first experiment generalise to full utterances and across different ASR systems.

2.7. Materials and methods

2.7.1. Participants

In Experiment 2, we included the same 21 Dutch-speaking mother-infant dyads from the larger cross-linguistic corpus of Dutch and Mandarin Chinese infant-directed speech (Han, 2019) that were used in the previous experiment.

2.7.2. Transcriptions

We used the same automatic transcriptions of the picture-book reading recordings that were described in the previous experiment generated by the open-source tool Kaldi-NL. Instead of only examining target words, however, we used the automatic transcriptions of the entire recordings of the picture-book reading sessions (M = 348 words per recording). To calculate WERs, we manually annotated all words in the IDS recordings. A research assistant was trained to manually correct and supplement the Kaldi-NL transcriptions. All words except for the occasional mentions of children's names were included in the annotation process. Children's names were also removed from the automatic transcriptions. The manual annotation procedure resulted in a gold standard IDS data set containing a total of 15,309 words. Out of this total data set, only 4.4% of the words were target words. The influence of target words in this experiment is thus minimal, as are their potential frequency effects on the outcomes.

For the comparison between two ASR systems, we also automatically transcribed the same IDS recordings using WhisperX (Bain et al., 2023) which provides improved accuracy and word-level timestamps using voice activity detection and forced phoneme alignment while using OpenAI's Whisper models (Radford et al., 2022). Whisper contains weakly supervised (or semi-supervised) cross-linguistic training models (i.e., audio paired with unvalidated transcripts from the Internet) which allows for a larger quantity of training data compared to supervised models. A larger quantity of training data could result in more robust models. The full data set comprises over 680,000 hours of training data of which 117,000 hours cover 96 other languages. When testing the largest Whisper model on the Fleurs data set, a mean WER of 4.4%

was found for English and a mean WER of 6.7% was found for Dutch (Radford et al., 2022). WhisperX also takes audio files as input (e.g., WAV) and generates text files containing all transcribed words and their corresponding timestamps as output which can be used in the evaluation procedure.

2.7.3. Evaluation procedure

To calculate WERs, we used the toolkit *sclite* version 2.10 from *SCTK* version 2.4.12 (SCTK, 2023) which is an open-source tool for scoring and evaluating the output of ASR systems. All reference and hypothesis transcription files were transformed to CTM format before being submitted to *sclite*. The tool calculates the WER in percentages for individual speakers by dividing the sum of word deletions, insertions, and substitutions by the total number of words in the human-labelled transcription. The higher the WER, the lower the accuracy of the transcription. We standardised the texts by making all words lowercase and removing all punctuation in the ASR output and the reference transcriptions. In addition, common abbreviations were spelled out in full (e.g., *'m => hem*, *z'n => zijn*).

2.7.4. Statistical analysis

In addition to reporting the overall WERs, a statistical analysis was carried out in *R* version 4.2.0 (R Core Team, 2022). We fitted a linear mixed-effects model using the lme4 package version 1.1-30 (Bates et al., 2015) with WERs for each speaker as continuous outcome variables. Each speaker has four WER scores: two generated by Kaldi-NL and two generated by WhisperX, one for each measurement point. We included the ASR system (Kaldi-NL or WhisperX) and Age (18 months and 24 months) as two dichotomous predictors to the model. We used dummy coding where Kaldi-NL and 18m were used as reference levels. We also added random intercepts for participants to the model. This allowed us to examine 1) whether WERs are affected by children's ages and 2) whether WERs are affected by the open-source ASR tool used to generate the transcriptions.

2.8. Results

Across both time points, Kaldi-NL had a mean WER of 40.12% (*SD* = 10.39). WhisperX had a mean WER of 22.49% (*SD* = 10.28). Table 2.4 presents descriptive results of WERs of Kaldi-NL and WhisperX for speech directed at 18-month-old and 24-month-old infants. There is a large difference in performance between Kaldi-NL and WhisperX, but only small differences in performance between the two measurement waves.

Table 2.4. Descriptive statistics of WERs (percentages) of transcriptions by Kaldi-NL and WhisperX for speech directed at 18-month-olds (18m) and 24-month-olds (24m).

	18m		24m	
ASR system	*M*	*SD*	*M*	*SD*
Kaldi-NL	41.97	11.64	38.27	8.86
WhisperX	21.84	11.52	23.14	9.11

First, we compared the model including the predictor ASR system to a null model without any predictors. The model including the predictor ASR system provided a significantly better fit to the data compared to the null model ($p < .001$). Then, we added the predictor Age to this model. The model including Age did not provide a significantly better fit to the data ($p = 0.627$). The ASR systems did not perform differently on IDS directed at 18-month-olds or 24-month-olds. The results of the final model are shown in Table 2.5. The results show that the ASR system WhisperX significantly reduced WERs with 17.63% (95% CI [-21.52, -13.74]) on average compared to Kaldi-NL.

Table 2.5. Results of the linear mixed-effects model (WER ~ ASR system + (1|subject)).

Predictor	Estimate	*SE*	*t*-value
(Intercept)	40.12	1.78	22.52
System (WhisperX)	-17.63	1.97	-8.95

2.9. Discussion

The first aim of Experiment 2 was to evaluate how open-source ASR systems perform at transcribing full utterances in Dutch IDS. The results show that Kaldi-NL does not perform very well at transcribing full utterances in IDS. The mean WER of 40.12% is much higher than previously reported by Tejedor-García et al. (2022) for the healthcare domain. They used Kaldi-NL to transcribe Dutch doctor-patient consultations and found a WER of 25.8%. This suggests that IDS is a difficult speech register to transcribe for this ASR system trained on ADS. We found that WhisperX performs significantly better. Although the mean WER of 22.49% is higher than the WER of 6.7% reported for the Fleurs data set using Whisper (Radford et al., 2022), WhisperX could be used as a starting point to facilitate the annotation process of IDS. This could drastically decrease the time that is currently needed for creating manual annotations – saving time and valuable resources.

The second aim of Experiment 2 was to assess whether the supervised models of Kaldi-NL perform differently at transcribing IDS than the much larger, semi-supervised models of Whisper. The two ASR systems differ vastly in the amount of training data. Where the models of Kaldi-NL are trained on approximately 900 hours of Dutch speech from television shows and lectures, Whisper is trained on 680,000 hours of cross-linguistic training data (of which 117,000 hours cover 96 other languages). We found that WhisperX performed significantly better than Kaldi-NL at transcribing full utterances in Dutch IDS. The large number of open-source ASR systems available can make it difficult for researchers to know which system best suits

their needs. If multiple open-source ASR tools are available in a language, we would advise researchers to assess which system performs better on their specific data set. The results of this experiment show that the type of ASR system can have a large influence on the accuracy of the transcriptions.

2.10. General discussion

In the current study, we aimed to assess to what extent the accuracy of open-source ASR tools is affected when transcribing maternal speech directed at 18-month-old and 24-month-old infants. This is the first study to examine the transcription accuracy of IDS using off-the-shelf ASR tools trained on large, (semi-)supervised ADS data sets. Currently, most researchers on IDS transcribe audio recordings manually from scratch. while a growing number of open-source ASR tools trained on large data sets are available cross-linguistically Although the manual procedure results in highly accurate transcriptions, it is labour-intensive which makes the data annotation process time-consuming and expensive. To date, no studies have examined whether researchers can successfully use off-the-shelf ASR tools trained on ADS for the annotation of IDS. Using automated tools can drastically decrease the time that is currently needed for manual transcriptions.

The results show that the open-source ASR system Kaldi-NL is less accurate when transcribing IDS compared to ADS. We found that the recognition accuracy of target words is decreased when they are produced in IDS compared to ADS, and we also found a negative effect of mean pitch. The difference in accuracy between the two speech registers was largest for speech directed at younger children. These results suggest that we first have to identify whether ASR tools can provide benefits before we start implementing them in the annotation process. A previous study found that WERs should be below 30% for automatic transcriptions to be beneficial to the annotation process (Gaur et al., 2016). Otherwise, it would be faster to annotate manually from scratch. Although we believe this limit can be different depending on the types of recognition errors or the specific research goals, the results of our study constitute evidence that the open-source ASR system WhisperX can transcribe Dutch

IDS at a more than sufficient accuracy. We would recommend researchers on IDS to compare the performance of multiple off-the-shelf ASR systems in case those are readily available in their language. The accuracy may differ depending on the characteristics of the training data or the data set being transcribed.

In Experiment 1, we found that the difference in recognition performance of Kaldi-NL at transcribing target words in IDS and ADS decreased over time. In Experiment 2, we did not find that Kaldi-NL or WhisperX performed differently across the two time points when comparing WERs of full sentences. There are two possible explanations for this result which are not mutually exclusive. First, the typical acoustic features of IDS could be more prominent in word-level acoustics compared to utterance-level acoustics. Previous work on this has found that when mothers are reading a picture book containing target words to their infants, mothers consistently positioned these words on exaggerated pitch peaks (Fernald & Mazzie, 1991). However, this result was not replicated in this Dutch IDS data set by Han et al. (2020). In Dutch IDS, the pitch of target words was similar to the pitch of utterances. Therefore, it is not likely that by examining target words, we have inflated the effects of the acoustics of the speech register IDS on recognition accuracy. An alternative explanation is that Kaldi-NL relies more on the acoustic model. Therefore, this system could be more affected by the acoustic differences of IDS which are more prominent when children are younger. For Kaldi-NL, we found a decrease in WER of 3.7% for speech directed at older children which is what we would expect if the system relies more heavily on the acoustic model. In contrast, we found an increase of 1.3% in WER for speech directed at older children for WhisperX. If WhisperX relies more on the language model, this could suggest that the performance of WhisperX is more impacted by the low frequency target words that were spoken to older children, rather than the acoustic differences across the two time points.

Future studies should examine whether we can improve the automatic annotation of IDS by applying front-end lowering of mean pitch of the speech recordings (see Gustafson & Sjölander, 2002 for the application of this method to children's speech).

This could be an efficient, cost-effective solution which can be easily applied by researchers studying different languages – provided a well-trained ASR system in their language exists. This solution, if successful, could be a simple method to create a small but significant improvement in recognition accuracy. Another approach that could be taken in future studies would be to train new language and/or acoustic models on IDS data. For this to work, the IDS data set must be large and general enough to be useful for application on new data sets.

2.11. Conclusions

In these experiments, we showed that open-source ASR systems can be used for the annotation of Dutch IDS. Although the performance decreases when transcribing IDS compared to ADS, the results are a promising start. Depending on the research goals, automatic transcriptions still need to be corrected by a human annotator. However, this correction process will take less time compared to transcribing the data from scratch. We additionally showed that the choice of ASR system has a large influence on the results. For our IDS data set, WhisperX performs significantly better than Kaldi-NL. This is the first study that assessed the accuracy of automatic transcriptions of (Dutch) IDS directed at children of different ages generated by different off-the-shelf ASR systems. While there is much room for improvement, the results show that automatic transcriptions provide a promising starting point for researchers who have to transcribe a large amount of speech directed at infants**.**

Acknowledgement and data availability

This work is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003) and by Utrecht University's Human-centered Artificial Intelligence focus area (HAI Summer 2021 Small Grant). Anonymised data frames and the *R* markdown script are available online: https://osf.io/jbg9t/.

References

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *The Journal of the Acoustical Society of America*, *141*(5), 3070–3078. https://doi.org/10.1121/1.4982246

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). *WhisperX: Time-accurate speech transcription of long-form audio* (arXiv:2303.00747). arXiv. http://arxiv.org/abs/2303.00747

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5–22. https://doi.org/10.1016/S0167-6393(00)00067-4

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

Burnham, D., Kalashnikova, M., Muawiyath, S., Cassidy, S., Estival, D. (2016). *Infant-directed speech research made easy: A database, some tools and a virtual laboratory*. Abstract and paper presented at the 43rd Experimental Psychology Conference, Melbourne, Australia.

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* [Computer program] (6.1.38). http://www.praat.org/

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

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1), 1–13.

Estes, K. G., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy: The Official Journal of the International Society on Infant Studies*, *18*(5). https://doi.org/10.1111/infa.12006

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior & Development*, *8*(2), 181–195. https://doi.org/10.1016/S0163-6383(85)80005-9

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*, 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, *20*(1), 104–113. https://doi.org/10.1037/0012-1649.20.1.104

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Gaur, Y., Lasecki, W. S., Metze, F., & Bigham, J. P. (2016). The effects of automatic speech recognition quality on human transcription latency. *Proceedings of the 13th International Web for All Conference*, 1–8. https://doi.org/10.1145/2899475.2899478

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, *52*(3), 181–200. https://doi.org/10.1016/j.specom.2009.10.001

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In D. E. Losada & J. M. Fernández-Luna (Eds.), *Advances in Information Retrieval* (pp. 345–359). Springer. https://doi.org/10.1007/978-3-540-31865-1_25

Gustafson, J., & Sjölander, K. (2002). Voice transformations for improving children's speech recognition in a publicly available dialogue system. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 297–300. https://doi.org/10.21437/ICSLP.2002-139

Han, M. (2019). *The role of prosodic input in word learning: A cross-linguistic investigation of Dutch and Mandarin Chinese infant-directed speech* [Dissertation, Utrecht University]. http://localhost/handle/1874/379614

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

Han, M., Jong, N. H. D., & 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

Hirschberg, J., Litman, D., & Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, *43*(1), 155–175. https://doi.org/10.1016/j.specom.2004.01.006

Johnson, E. K., Lahey, M., Ernestus, M., & Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *The Journal of the Acoustical Society of America*, *134*(6), EL534–EL540. https://doi.org/10.1121/1.4828977

Kawahara, T., Nanjo, H., Shinozaki, T., & Furui, S. (2003). Benchmark test for speech recognition using the corpus of spontaneous Japanese. *Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, paper TMO4.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238–2246. https://doi.org/10.1121/1.1869172

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, *24*(4), 372–392. https://doi.org/10.1016/S0163-6383(02)00086-3

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. https://doi.org/10.1126/science.277.5326.684

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

Mohamed, A., Hinton, G., & Penn, G. (2012). Understanding how Deep Belief Networks perform acoustic modelling. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4273–4276. https://doi.org/10.1109/ICASSP.2012.6288863

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first evaluation. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. LREC 2000, Athens, Greece.

Peddinti, V., Wang, Y., Povey, D., & Khudanpur, S. (2018). Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters*, *25*(3), 373–377. https://doi.org/10.1109/LSP.2017.2723507

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Veselý, K. (2011). *The Kaldi Speech Recognition Toolkit*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision* (arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Shinozaki, T., Hori, C., & Furui, S. (2001). Towards automatic transcription of spontaneous presentations. *EUROSPEECH-2001*, 491–494.

Shinozaki, T., Ostendorf, M., & Atlas, L. (2009). Characteristics of speaking style and implications for speech recognition. *The Journal of the Acoustical Society of America*, *126*(3), 1500–1510. https://doi.org/10.1121/1.3183593

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*, 654–666. https://doi.org/10.1080/15250000903263973

Sjons, J., Hörberg, T., Östling, R., & Bjerva, J. (2017). Articulation rate in Swedish child-directed speech increases as a function of the age of the child even when surprisal is controlled for. *ArXiv:1706.03216 [Cs]*. http://arxiv.org/abs/1706.03216

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

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

Spinelli, M., Fasolo, M., & Mesman, J. (2017). Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Developmental Review*, *44*, 1–18. https://doi.org/10.1016/j.dr.2016.12.001

Tejedor-García, C., van der Molen, B., van den Heuvel, H., van Hessen, A., & Pieters, T. (2022). Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1032–1039. https://aclanthology.org/2022.lrec-1.110

Yilmaz, E. & van Gompel, M. (2020). *Automatic Transcription of Dutch Speech Recordings* [ASR tool] (0.5.0). https://webservices.cls.ru.nl/asr_nl

Zangl, R., Klarman, L., Thal, D., Fernald, A., & Bates, E. (2005). Dynamics of Word Comprehension in Infancy: Developments in Timing, Accuracy, and Resistance to Acoustic Degradation. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, *6*(2), 179–208. https://doi.org/10.1207/s15327647jcd0602_2

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., Zhou, Z., Li, B., Ma, M., Chan, W., Yu, J., Wang, Y., Cao, L., Sim, K. C., Ramabhadran, B., … Wu, Y. (2022). BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1519–1532. https://doi.org/10.1109/JSTSP.2022.3182537

SCTK, the NIST Scoring Toolkit. (2023). [C]. National Institute of Standards and Technology. https://github.com/usnistgov/SCTK (Original work published 2016)

Chapter 3

Caregiver reports of Dutch children's vocabularies: Effects on vocabulary size are age-specific and task-specific

Abstract

Limited studies have examined demographic differences in vocabulary over time, while there are questions regarding the onset and stability of these effects on caregiver reports versus lab-administered vocabulary tasks. In this longitudinal study, we included over 300 Dutch children from the YOUth cohort. Caregivers filled out adapted versions of the N-CDIs when children were around 10 months (measuring word comprehension, word production, and gestures) and around 3 years of age (measuring word production). In Part 1, we assessed the validity of the N_{YOUth}-CDIs. They show concurrent and predictive validity – also with a lab-administered vocabulary task. In Part 2, we examined the longitudinal effects of predictors on N_{YOUth}-CDIs and the lab-administered task. Although we found that children's gender, maternal education, and multilingualism explained some variance in children's vocabularies, the effects were age-specific and task-specific. Examining predictors in longitudinal samples helps build a comprehensive understanding of the influences on early vocabulary development.

This chapter has been submitted to a journal:

van der Klis, A., Junge, C., Adriaans, F., & Kager, R. (submitted). Caregiver reports of Dutch children's vocabularies: Effects on vocabulary size are age-specific and task-specific.

3.1. Introduction

The MacArthur–Bates Communicative Development Inventories (CDIs), have been adapted into many languages, including Flemish Dutch (N-CDIs; Zink & Lejaegere, 2002). Such adaptations have two advantages: they allow testing similarities in cross-linguistic patterns (e.g., Frank et al., 2021) as well as offering a unified tool suitable for a wide range of languages to examine individual variation in vocabulary development (e.g., Cristia et al., 2014). Indeed, CDIs are employed in many longitudinal cohorts centring on capturing individual variation in early development (e.g., Kidd et al., 2018; Peter et al., 2019; Reilly et al., 2009; Verhoef et al., 2021). In the Netherlands, the YOUth cohort study follows thousands of children whose caregivers fill out adapted versions of N-CDIs (hereafter N_{YOUth}-CDIs) when the child is around 10 months old (measuring word comprehension, word production, and gestures) and around 3 years old (measuring word production) (Onland-Moret et al., 2020). The N_{YOUth}-CDIs were adapted in several ways. First, we used short forms to save time as caregivers in the YOUth cohort study have to fill out a broad range of questionnaires. The norming study reports a high correlation between short forms and full-length forms (Zink & Lejaegere, 2003). We included the gesture scales from the full-length N-CDI which could be a more relevant scale for capturing individual variation across 10-month-old infants. Second, we changed or removed Flemish Dutch words to better fit the dialect of Dutch spoken in the Netherlands. Lastly, we combined the N-CDI 2 and N-CDI 3 to better accommodate the age range of children included in the second measurement wave. The adaptations require us to validate the N_{YOUth}-CDIs before we continue to use them to examine individual variability in children's vocabularies.

Many studies have shown that CDIs are reliable and valid tools for measuring early vocabulary across a wide range of participants (Feldman et al., 2005; Fenson et al., 2007; Frank et al., 2021). There are advantages compared to using naturalistic speech samples or lab-administered tasks to measure children's vocabularies. Administering CDIs is a standardised, fast, and cost-effective approach that does not require trained

lab assistants, lab visits, or the labour-intensive transcription of speech. This allows for larger sample sizes which could be beneficial, especially when examining environmental influences on vocabulary development. In addition, CDIs are useful additions to laboratory samples which provide snapshots of children's language use. However, caregivers are prone to several reporting biases. Caregivers with lower levels of education tend to report larger vocabularies than caregivers with higher education, especially for word comprehension in infants (Bavin et al., 2008; Feldman et al., 2000; Fenson et al., 2007; Reese & Read, 2000), while studies typically report positive effects of maternal education on CDIs administered with toddlers (Feldman et al., 2000; Fenson et al., 2007, but cf. Reese & Read, 2000; Kuvač-Kraljević et al., 2021). Caregivers of lower education could have more liberal criteria for determining word comprehension compared to caregivers with higher education, and/or caregivers tend to overreport for infants with smaller vocabularies because they think larger vocabularies are more desirable (for discussions, see Feldman et al., 2000; Tomasello & Mervin, 1994). The effects of environmental predictors of variation in children's vocabularies vary based on differences in sample characteristics and/or the vocabulary measure being used.

In Part 1, we set out to examine the reliability and validity of the N_{YOUth}-CDIs for Dutch children around 10 months of age (Wave 1) and around 3 years of age (Wave 2). First, we assessed correlations across the N_{YOUth}-CDI scales at both ages; then we examined their concurrent and predictive correlations with the Peabody Picture Vocabulary Test (PPVT) administered during Wave 2. In Part 2, we used this large, longitudinal sample of Dutch children to examine whether the effects of key predictors of variation in children's vocabularies are age-specific and task-specific. Previous studies have identified key predictors of variation in children's early vocabularies – including maternal education, children's gender, gestational age, birthweight, and multilingualism – but there are uncertainties regarding their effects on different vocabulary outcomes across development. Examining the effects of predictors in a large, longitudinal sample across multiple vocabulary measures is an important step in validating the generalisability and stability of the predictors. This

helps build a more comprehensive understanding of the influences on early vocabulary development.

3.2. Part 1: Validity and reliability

3.2.1. Early measures predict later vocabulary

Infants start developing their receptive vocabularies (i.e., language comprehension) before they learn how to speak. By 6–9 months of age, experimental research shows that English-learning infants already understand the meaning of some common nouns (Bergelson & Swingley, 2012). Children's earliest perceived *and* produced words are similar across languages and typically include important family members (e.g., "mommy"), social routines (e.g., "peekaboo"), and sound effects ("broom broom") (Frank et al., 2021). At the same time, vocabulary development is characterised by large differences in the onset and development rates of individual children within and across languages. These early differences have an impact on children's later vocabularies. Receptive vocabulary measured around 12 months of age positively correlates with Dutch children's receptive (r = .48) and productive (r = .29) vocabularies at 24 months (Zink & Lejaegere, 2002). The perception and production of words is not the only aspect of early vocabulary development. As an earlier means of communication, infants start using gestures. The full N-CDI-Words and Gestures (WG) not only asks about the comprehension and production of words but also whether the child is already capable of making certain actions and gestures, including the first communicative gestures (e.g., index-finger pointing), games and routines (e.g., playing peekaboo), actions with objects (e.g., eating with a spoon or fork), and pretending to be a caregiver (e.g., pretending to feed a doll) (Zink & Lejaegere, 2002).

Many studies have highlighted the relationship between infants' gestures and their later vocabularies (e.g., Brooks & Meltzoff, 2008; Colonnesi et al., 2010; McGillion et al., 2017; Rowe & Goldin-Meadow, 2009). For example, Brooks and Meltzoff (2008) have shown that infants who pointed had faster vocabulary growth during the second year of life compared to non-pointers. We found that infants' pointing gestures

elicit more verbal responses from caregivers compared to infants' prelinguistic vocalisations or other types of gestures (see Chapter 4). The relationship between infants' gestures and vocabulary development could be mediated through caregivers' responses (Olson & Masur, 2015). Many previous studies measured infant gestures during observations in the lab or at home (e.g., Brooks & Meltzoff, 2008; McGillion et al., 2017; Rowe & Goldin-Meadow, 2009) which is time-consuming work to record and annotate by hand. The gesture scales included in the CDI-WG correlate with children's later vocabularies (e.g., Cadime et al., 2017; Fenson et al., 1994; Kuvač Kraljević et al., 2014; Sansavini et al., 2010). For children aged 8–16 months, the gesture scales capture more individual variation than word production or word comprehension as shown by the average item difficulty (i.e., percentage of participants who respond correctly). At 10 months of age, most test items for word comprehension and word production are understood or produced by less than 40% of infants, suggesting that all items have high difficulty for this age group. For gestures, there are more items (11%) that are understood by at least 40% of infants, suggesting that the items included in the gesture scale capture more individual variation within this age group (Zink & Lejaegere, 2002). This makes it important to include gesture scales in caregiver reports when the goal is to capture individual differences across infants in the first year of life.

3.2.2. Reliability and validity of caregiver reports

Caregiver reports of infants' vocabularies show high reliability and moderate to strong concurrent and predictive validity (e.g., Bates et al., 1995; Fenson et al., 2007; O' Toole & Fletcher, 2010; Pan et al., 2004; Reese & Read, 2000; Zink & Lejaegere, 2002). Previous studies assessing the validity of CDIs calculated concurrent correlations across CDI scales. For the N-CDI 1 shortlist containing 103 items, receptive vocabulary correlates positively with productive vocabulary (r = .62) (Zink & Lejaegere, 2003). For the full-length N-CDI-WG, gestures correlate positively with word production (r = .41) and word comprehension (r = .57) (Zink & Lejaegere, 2002). However, as noted by Tomasello and Mervis (1994), correlations between scales also measure the degree to which caregivers were consistent in their judgments.

Hence, we should also assess test validity by assessing the concurrent and predictive relationships between caregiver reports and vocabulary measures obtained on another standardised task, such as the Peabody Picture Vocabulary Test (PPVT) which is a lab-administered task measuring receptive vocabulary skills in children aged 2.3 years and older (Dunn & Dunn, 1997). Previous studies found moderate to strong concurrent (r = .41 – .50) and predictive (r = .32 – .48) correlations between CDIs and PPVT scores assessed within a 7-month to an almost 2-year time frame in between measurements (e.g., Feldman et al., 2005; Pan et al., 2004; Reese & Read, 2000). This supports the reliability and validity of using caregiver reports to measure infants' early vocabularies.

3.2.3. Research aim

YOUth is an ongoing, longitudinal cohort study following Dutch children prenatally up to early childhood (Onland-Moret et al., 2020). Around 10 months of age, we collect the N$_{YOUth}$-CDI 1 including three vocabulary scales: vocabulary production, vocabulary comprehension, and gestures. When children are around 3 years of age, we collect the N$_{YOUth}$-CDI 2 including vocabulary production. During this wave, we also collect the PPVT-III-NL receptive vocabulary task in the lab. Due to adaptations to the N-CDIs, we must assess the measurement quality before we continue to use them to study individual variation in children's vocabularies. The present study sets out to examine the validity and reliability of the N$_{YOUth}$-CDIs to assess Dutch children's vocabularies. We examined the internal consistencies of N$_{YOUth}$-CDI scales as indicators of test reliability and the concurrent validity across N$_{YOUth}$-CDI 1 scales, the concurrent validity between the N$_{YOUth}$-CDI 2 and PPVT-III-NL, and the predictive validity of the N$_{YOUth}$-CDI 1 scales and PPVT-III-NL as indicators of test validity. Concurrent or predictive relationships between N$_{YOUth}$-CDIs and a standardised, lab-administered task such as the PPVT-III-NL provides us with solid evidence of test validity.

3.3. Methods

3.3.1. Participants

The data for this study are derived from the YOUth cohort study which involves repeated measurements at regular intervals. From the cohort, 444 Dutch infants around 10 months of age (230 girls, age *M* = 10.6 months, range = 9.0 – 13.1 months, *SD* = 0.9) (hereafter Wave 1) were included in this study. These were all the children in the YOUth cohort study who had participated in the next wave by March 2022. During this wave, the same children were on average 3.4 years of age (range = 2.0 – 6.0 years, *SD* = 0.8) (hereafter Wave 2). There were approximately one to five years (*M* = 2.5, *SD* = 0.8) in between measurement waves, randomly varying per participant. We followed the Code of Ethics of the World Medical Association (Declaration of Helsinki), and all caregivers signed informed consent prior to participating. During Wave 1, children received a Miffy picture book for their participation. During Wave 2, children received a frog umbrella.

3.3.2. Materials and procedure

N$_{YOUth}$-CDIs

We administered the N$_{YOUth}$-CDI 1 — measuring vocabulary production, vocabulary comprehension, and gestures — during Wave 1. The N$_{YOUth}$-CDI 1 contains the short form of words (Zink & Lejaegere, 2003). We used the short form because it contains only 103 compared to 434 items, which makes this form far less time-consuming to complete. This was important since caregivers already have to fill out a broad range of questionnaires in the YOUth cohort study. Caregivers were asked to check for each item whether their child *understands* or *speaks* the word — also when the child produces synonyms or pronunciation errors. In the N$_{YOUth}$-CDI 1, we replaced or removed 12 typical Flemish words with synonyms that are more common in Standard Dutch spoken in the Netherlands (e.g., we removed *mantel* from *jas(je) / mantel* ("coat")). Given the important role of gestures in early vocabulary development, we included the list containing 65 gestures and actions from the full-length N-CDI-WG (Zink & Lejaegere, 2002) which is usually not included in the short forms. This scale

contains "early gestures" including the first communicative gestures (e.g., pointing) and games and routines (e.g., playing peekaboo) and "late gestures" including actions with objects (e.g., eating with a spoon or fork) and pretending to be a caregiver (e.g., pretending to feed a doll). The NYOUth-CDIs were emailed to the primary caregiver. The NYOUth-CDIs are fully digitised so caregivers could fill them out online. We scored the lists following the instructions of the manuals (Zink & Lejaegere, 2002, 2003).

The NYOUth-CDI 2 is a contraction of the short forms N-CDI 2A (16-30 months) and N-CDI 3 (30-37 months) (Zink & Lejaegere, 2003). This was necessary because there was only one measurement wave (Wave 2) during the toddler and preschool years in the YOUth cohort study. The contraction resulted in a total number of 207 vocabulary items after removing the overlapping ones. Caregivers are asked to check the items that the child *speaks* – also in case the child produces synonyms or pronunciation errors. In the NYOUth-CDI 2, we also replaced or removed 26 typical Flemish words with similar words that are more common in Standard Dutch spoken in the Netherlands (e.g., *bank* instead of *zetel/sofa* ("couch")). The CDIs for toddlers (including adaptations in other languages) do not measure vocabulary comprehension or gestures anymore. Most toddlers and older children have already acquired all the gestures resulting in a ceiling effect. Children of this age group are also old enough to participate in a lab-administered task of vocabulary comprehension. Caregivers were instructed to fill the NYOUth-CDI 2 out within four weeks after the administration of the PPVT-III-NL in the lab during Wave 2.

Peabody Picture Vocabulary Task

During Wave 2, we also administered the third version of the Dutch Peabody Picture Vocabulary Task (PPVT-III-NL) which is a lab-administered task of receptive vocabulary (Schlichting, 2005). The task measures whether a person can match a spoken word to one of the four pictures (i.e., multiple choice). It is designed as a behavioural task in which the participant points to one of the images and the experimenter produces the target words and scores manually. For the YOUth cohort

study, we developed a computerised version of the PPVT-III-NL. The experimenter runs a script on a computer with a touch screen where children are provided with recordings of the test items and four pictures on the screen. This controls for differences in speaker pronunciations and minimises the role of the experimenter. Children can use the touch screen to select one of the pictures after the target item has been presented. During the task, words become increasingly more complex. The PPVT-III-NL has a total of 204 items, divided into 17 sets of 12 items. The task terminates when the child makes nine or more errors in one set ("final set") (see Schlichting, 2005). The programme automatically subtracts the number of errors from the maximum score (which is the number of the final set * 12 items), resulting in the child's raw score. During the task, the child's caregiver was present in the back of the room out of the child's view. Caregivers were explicitly instructed not to help or communicate with the child.

3.3.3. Coding and analyses

All analyses were carried out in *R* version 4.2.0 (R Core Team, 2022). For the NYOUth-CDI 1, we calculated "vocabulary production" by summing all vocabulary items for which caregivers ticked the box *speaks*, "vocabulary comprehension" by summing all vocabulary items for which caregivers ticked the box *understands* or *speaks*, and "total gestures" by summing all *yes*, *sometimes*, and *often* responses on the gesture scale. Gestures can be subdivided into two categories: "early gestures" and "late gestures" (Zink & Lejaegere, 2002). The sum of both scales results in the score "total gestures". We used these raw scores to analyse the data. For the NYOUth-CDI 2, we calculated "vocabulary production" by summing all items that were marked by the caregivers indicating that the child produces the word. For the PPVT-III-NL, we obtained "vocabulary comprehension" through the raw scores which were automatically calculated by the computer script, and we converted raw scores to norm scores based on children's ages on the day of the PPVT-III-NL administration (see Schlichting, 2005).

First, to determine the internal consistencies of the scales, we calculated Cronbach's alpha (α) using the package *ltm* version 1.2-0 (Rizopoulos, 2006). Cronbach's alpha estimates the average of all possible split-half correlations for test items. When comparing groups, α-values > 0.70 are satisfactory, while for clinical applications α-values > 0.90 are desired (Bland & Altman, 1997). Then, we used correlation tests to determine correlations between word production, word comprehension, and gestures. We used Pearson's correlations to determine any relations between word comprehension and gestures, and we used the non-parametric Spearman's correlation test for word production to account for the non-normal distribution of word production scores. We used partial correlations correcting for the varying age gap between Wave 1 and Wave 2 using the package *ppcor* version 1.1 (Kim, 2015) when calculating predictive validity.

3.4. Results

3.4.1. Descriptive statistics

We included 444 participants from the YOUth cohort study. During Wave 1, 338 of these participants completed the N$_{YOUth}$-CDI 1. There was one participant who did not complete the gestures list; this participant is only excluded from analyses involving gestures. During Wave 2, we had to exclude four participants from the PPVT-III-NL because the children did not participate (n = 2) or the test day had ended prematurely before administering the PPVT-III-NL (n = 2) resulting in no data. We excluded an additional 11 children from any analyses involving the PPVT-III-NL because they did not finish the task, resulting in 429 participants. There were 303 participants whose caretakers completed the N$_{YOUth}$-CDI 2 for Wave 2. The descriptive results of the vocabulary tests are presented in Table 3.1. The high standard deviations indicate that vocabulary scores are spread out over a wide range, revealing a large amount of individual variability.

Table 3.1. Descriptive results including the mean (*M*) and standard deviation (*SD*) of the different vocabulary measures.

Wave	*n*	Comprehension[a] *M* (*SD*)	Production *M* (*SD*)	Gestures *M* (*SD*)
1	337-338	36.75 (21.79)	2.86 (4.22)	18.47 (7.83)
2	303-429	52.82 (18.37)	173.4 (36.92)	

[a] At Wave 1, vocabulary comprehension is measured with the N_{YOUth}-CDI 1. At Wave 2, vocabulary comprehension is measured with the PPVT-III-NL (raw scores).

3.4.2. Internal consistency

First, we examined whether the N_{YOUth}-CDIs showed internal consistency. For the N_{YOUth}-CDI 1, we calculated Cronbach's alpha separately for comprehension (α = .97), production (α = .91), and gestures (α = .89) which represents the consistency of items within each scale. We also calculated Cronbach's alpha for the N_{YOUth}-CDI 2 word production (α = .99) indicating that the items on the scale measured the same construct. Overall, this indicates the caregiver reports of children's vocabularies show excellent internal consistency.

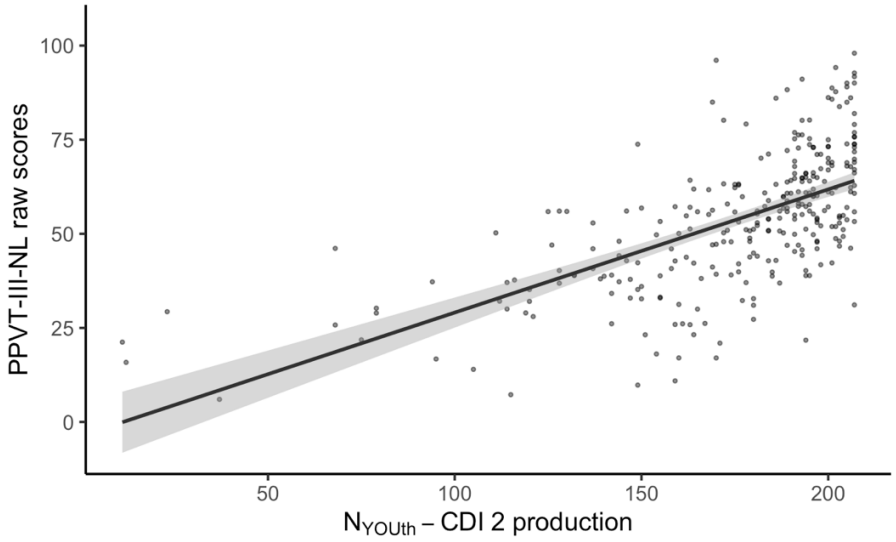
3.4.3. Concurrent validity

The results of the correlation tests indicate that for the N_{YOUth}-CDI 1, comprehension was positively correlated with both production, $r_s(336) = .50$, $p < .001$ and gestures, $r(335) = .65$, $p < .001$. Production was also correlated positively with gestures, $r_s(335) = .47$, $p < .001$.

We also examined whether vocabulary production obtained by the N_{YOUth}-CDI 2 shows a concurrent relationship with vocabulary comprehension measured by the lab-administered PPVT-III-NL at the same time. The correlation test indicates there is a strong, positive correlation between N_{YOUth}-CDI 2 production and concurrent PPVT-

III-NL comprehension scores, $r_s(292) = .65$, $p < .001$. The relationship is depicted in Figure 3.1.

Figure 3.1. The concurrent relationship between N_{YOUth}-CDI 2 production and PPVT-III-NL comprehension including a linear regression line with a 95% confidence interval.



3.4.4. Predictive validity

The last step to assess test validity was to calculate the predictive validity of the N_{YOUth}-CDI 1. We examined whether vocabulary production, vocabulary comprehension, and gestures measured at Wave 1 were correlated with N_{YOUth}-CDI 2 production and PPVT-III-NL comprehension measured at Wave 2. In total, 266 participants completed the N_{YOUth}-CDIs during Wave 1 and Wave 2, and 325 participants completed both the N_{YOUth}-CDI 1 at Wave 1 and the PPVT-III-NL at Wave 2. We ran partial correlations correcting for the varying time interval between the two waves. We therefore used PPVT-III-NL raw scores which are not yet

corrected for age. The results of all (partial) correlation tests are summarised in Table 3.2.

Table 3.2. Partial correlation table (controlling for the time gap between Wave 1 and Wave 2 for predictive relations) showing the links between the different NYOUth-CDI scales at Wave 1 and Wave 2 and the PPVT-III-NL at Wave 2.

	1	2	3	4
1. NYOUth-CDI 1 Comprehension	–			
2. NYOUth-CDI 1 Production	.50***	–		
3. NYOUth-CDI 1 Gestures	.65***	.47***	–	
4. NYOUth-CDI 2 Production	.31***	.17**	.15*	–
5. PPVT-III-NL Comprehension	.08	.08	.15**	.65***

The results show that all measures of the NYOUth-CDI 1 were positively correlated with later NYOUth-CDI 2 production scores. Overall, the strengths of the correlations were weak to moderate. We also found that comprehension at Wave 2 (i.e., PPVT-III-NL) only correlated with the gesture scale in Wave 1.

3.4.5. Early and late gestures

We found that the gesture scale correlates positively with both later NYOUth-CDI 2 production and PPVT-III-NL receptive vocabulary. As an exploratory analysis, we next examined whether early and late gestures are differentially related to these vocabulary outcomes by performing Spearman's partial rank-order correlations between early and late gestures and the two vocabulary outcomes separately. We again corrected for the varying time interval between the two waves. The results show that early gestures positively correlated with both PPVT-III-NL comprehension, $r(321) = .14$, $p < .05$ and NYOUth-CDI 2 production, $r(262) = .19$, $p < .01$, while late gestures only positively correlated with PPVT-III-NL comprehension, $r(321) = .17$, $p < .01$, but not with NYOUth-CDI 2 production, $r(262) = .11$, $p = .07$.

3.5. Discussion

3.5.1. Reliability and relations across scales

The results of our study show that the NYOUth-CDIs are valid measures for obtaining vocabulary data of infants and toddlers. The results show that the NYOUth-CDIs have good to excellent internal consistency, indicating that items in each scale measure a similar construct. The moderate to strong correlations between the separate components of the NYOUth-CDI 1 additionally suggest that the separate scales (i.e., vocabulary production, vocabulary comprehension, and gestures) are valid. Overall, the correlation coefficients of concurrent validity were slightly lower than the ones reported by Zink and Lejaegere (2002, 2003). Given the fact that we only examined infants around 10 months of age and concurrent relationships become stronger when children grow older (Zink & Lejaegere, 2002), this was to be expected. For production data, we found a floor effect because most infants produced none or only a few words during Wave 1. The results of our study suggest that despite the floor effect, production is significantly correlated with comprehension and gestures at this early age, and it is a weak predictor of later production measured by the NYOUth-CDI 2. We also found that NYOUth-CDI 1 comprehension and gestures were weakly to moderately related to later NYOUth-CDI 2 production. Yet, relations across CDI scales also measure the consistency with which caregivers fill out the reports (see Tomasello & Mervis, 1994). The results on different scales could each be influenced by similar reporting biases. Therefore, it is also vital to establish validity by examining the relations of CDIs with vocabulary measures obtained through another standardised task, such as the PPVT-III-NL.

3.5.2. Relations with the PPVT-III-NL

First, we found that the NYOUth-CDI 2 strongly correlated with the PPVT-III-NL, even though the latter task measured vocabulary comprehension rather than vocabulary production. This concurrent relationship confirms that the NYOUth-CDI 2 is a valid measure for establishing toddlers' vocabulary sizes. While some children in our sample were too old for the N-CDI 3 where the NYOUth-CDI 2 was partially based on

(above 37 months) or too young for the PPVT-III-NL (below 27 months), we found a stronger correlation than previously reported for the American English CDIs and the PPVT (Feldman et al., 2005). We administrated the PPVT-III-NL task in a touch-screen adaptation, presenting recordings of test items, and automatic scoring of the N$_{YOUth}$-CDIs and PPVT-III-NL which could have minimised any influences of the experimenter or human errors on vocabulary outcomes. This could have enhanced the validity of the task. However, given this ceiling effect in N$_{YOUth}$-CDI 2 production, we would not recommend using only the N$_{YOUth}$-CDI 2 for typically developing children older than 37 months if the goal is to capture individual variability within this age group. For older children, the PPVT-III-NL is a more suitable task to capture all variability that we typically find in vocabulary data.

We also examined the predictive validity of N$_{YOUth}$-CDI 1 scales for later PPVT-III-NL scores. We found that, although weakly, gestures positively correlated with the PPVT-III-NL. Previous studies have shown that differences across infants in gestures explain differences in their later vocabularies (e.g., Brooks & Meltzoff, 2008; Rowe & Goldin-Meadow, 2009). The positive relationship found in our study could be driven by the presence of specific gestures in infants' repertoires, such as index-finger pointing, which elicit more relevant verbal responses from caregivers (see Chapter 4). Nevertheless, the results of the exploratory analysis revealed that early gestures (including deictic gestures) *and* late gestures (including actions with objects) are differentially related to children's later vocabularies. Around 10 months of age, children's early gestures predicted their later word comprehension and word production, while late gestures only predicted later word comprehension. This is in line with earlier findings by Sansavini et al. (2010) who found that concurrently, "actions with objects" only showed a tight relationship with word comprehension but not with word production. Our findings add the observation that also late gestures (including "actions with objects") measured around 10 months can predict children's word comprehension – but not word production – several years later. When analysing early and late gestures together, studies typically find stronger correlations with

receptive vocabularies, which in turn lead to larger expressive vocabularies (e.g., Goldin-Meadow et al., 2007; Kuvač Kraljević et al., 2014).

For infants around 10 months of age, both the early gestures and late gestures scales can capture sufficient individual variation which can predict children's vocabulary outcomes. We hypothesise that the predictive values of the different gesture scales can change throughout children's development. While early gestures seem to have predictive validity for infants around 10 months, we can assume that these scales fail to capture enough individual variation across infants past a certain point in development. For older infants, it could become more informative to examine individual differences in late gestures which typically emerge in the second year of life. This idea is corroborated in an earlier study by Kuvač Kraljević et al. (2014) who found that for younger infants (8–12 months), deictic gestures, object gestures, and gestural routines correlated with word production. However, for older infants (13–16 months), deictic gestures did not correlate with word production. One-year-olds may not show sufficient variation in deictic gestures anymore due to a ceiling effect. This suggests that deictic gestures are most useful when measured in the second semester of life. This hypothesis can be tested in future studies sampling the gesture subscales across infants at multiple time points.

In short, we conclude that the N_{YOUth}-CDIs are reliable for infants around 10 months of age and children around 3 years of age. If the goal is to capture individual variation across infants, we recommend including the gesture scale. Given the fact that we first6 measured vocabulary at a very young age, and that the large varying time interval between the two measurement waves (one to five years), we consider the predictive relationships to be good.

3.6. Part 2: Demographic factors

Children show large individual differences in their vocabularies around 10 months and 3 years of age as shown in Table 3.1. The variation can partially be explained by genetics, but environmental variables most significantly influence children's

vocabularies in the early years (for a review, see Kidd & Donnelly, 2020). In this next part, we will examine widely reported demographic influences — maternal education, children's gender, gestational age, birthweight, and multilingualism — on children's vocabulary outcomes in this large, longitudinal sample of Dutch children.

3.6.1. Maternal education

Maternal education is often used as a proxy for socio-economic status (SES). Mothers with a higher educational background produce higher quantity (i.e., they generally speak more) and quality (i.e., they use more diverse language) of speech towards their children, mediating the positive relationship between maternal SES and children's language development (e.g., Hoff, 2003; Huttenlocher et al., 2010). Previous studies often reported positive effects of maternal education on children's vocabularies measured by CDIs for toddlers (e.g., Feldman et al., 2000; Fenson et al., 2007, but cf. Reese & Read, 2000; Kuvač-Kraljević et al., 2021). However, studies employing CDIs have frequently observed negative correlations between maternal education and children's vocabularies during infancy (e.g., Bavin et al., 2008; Feldman et al., 2000; Reese & Read, 2000). This early negative effect of maternal education on CDIs is likely driven by a caregiver reporting bias: A negative effect of SES is more often reported for vocabulary comprehension which requires more interpretation by the caregiver than vocabulary production, although a negative effect is sometimes reported for production as well (Bavin et al., 2008; Reese & Read, 2000). In contrast, studies rarely report a negative effect of SES on the gesture scale (Bavin et al., 2008; Feldman et al., 2000; Rowland et al., 2022). Determining whether a child produces a word or gesture does not require the caregiver to draw inferences about the child's understanding. In addition, there are fewer expectations from caregivers surrounding children's gesture development compared to their vocabulary development. On the one hand, caregivers could believe that larger vocabularies are more desirable – leading to over-reporting of their infants' vocabularies, or simply because some caregivers have more liberal criteria for word comprehension than others (see Feldman et al., 2000; Tomasello & Mervis, 1994 for discussions). On the other hand, caregivers may underestimate what their children already know when their children

do not produce many words yet (see Houston-Price et al., 2007). These findings make it relevant to study the effects of maternal education in large, longitudinal samples throughout the first years of development on a variety of vocabulary measures.

3.6.2. Children's gender

Many studies have identified small effects of children's gender. More specifically, girls tend to outperform boys on many vocabulary scales (e.g., Eriksson et al., 2012; Feldman et al., 2005; Frank et al., 2021; Reese & Read, 2000; Reilly et al., 2009; Zink & Lejaegere, 2002, but cf. Bavin et al., 2008). Simonsen et al. (2014) showed that boys are characterised by a less steep increase in receptive vocabulary growth than girls — at least until 20 months of age. Feldman et al. (2000) examined over 2,000 American English children using CDIs and reported lower scores for boys in vocabulary production and vocabulary comprehension across children aged 10–13 months. These differences persisted for older children, except for vocabulary comprehension. Girls have also been found to have larger gesture repertoires than boys based on CDIs (Feldman et al., 2000; Germain et al., 2022; Simonsen et al., 2014; Zink & Lejaegere, 2002). These studies suggest that overall, girls have faster developmental trajectories than boys. In contrast, previous studies using naturalistic speech samples or lab-administered tasks of children's receptive vocabularies typically do not report gender differences in diverse samples (e.g., Huttenlocher et al., 2010; Pan et al., 2004; Washington & Craig, 1999), although these findings are inconsistent, particularly for children's expressive language skills where girls tend to outperform boys (e.g., Bornstein et al., 1998; Frank et al., 2021; Qi et al., 2003). The effect of gender could be small and variable across children's ages and vocabulary measures, causing inconsistent results across studies.

3.6.3. Gestational duration and birthweight

Some studies suggest that preterm children are at a larger risk of having smaller vocabularies than full-term children (e.g., Foster-Cohen et al., 2007; Guarini et al., 2009; Sansavini et al., 2011, but cf. Ogneva & Pérez-Pereira, 2023). There may be negative effects only in extremely or very preterm children. Kern and Gayraud (2007)

found that very preterm (28–32 weeks) and extremely preterm (under 28 weeks) children had smaller vocabulary sizes based on CDIs than moderately preterm (33–36 weeks) and full-term children when they were assessed at 24–26 months of age. However, Pérez-Pereira and Cruz (2018) found that gestational age did not affect vocabulary growth in a sample of low-risk preterm children with a wide range of gestational ages and birthweights without other medical complications. Still, a meta-analysis showed that very preterm (under 32 weeks) and/or very low birthweight (under 1500 g) children have persistent language delays (Barre et al., 2011). Moreover, differences between preterm and full-term children in gestural and lexical development become increasingly more evident during the first two years of life (Sansavini et al., 2011; van Baar et al., 2006). Previous studies have not concurrently examined the effects of gestational duration and birthweight, and it remains a question whether these factors influence children's vocabulary development in a non-clinical sample. It also remains largely understudied whether vocabulary differences between preterm and full-term children are apparent during the first year of life. Therefore, it is relevant to study the effects of gestational age and birthweight in a large, longitudinal sample starting from infancy.

3.6.4. Multilingualism

In many studies examining test validity and factors influencing children's vocabularies using the CDIs, multilingual children are excluded. CDI norming samples also typically exclude multilingual children, while being multilingual is the norm in most places across the world. Therefore, it is important to assess how multilingualism affects children's performance on widely used vocabulary tasks. When assessing only one language, multilingual children have smaller vocabularies than their monolingual peers (Blom et al., 2020; De Houwer et al., 2014; Hoff et al., 2012). De Houwer et al. (2014) showed using CDIs that monolingual toddlers knew more Dutch words than bilingual toddlers (20 months), but both groups understood and produced the same number of lexicalised meanings. They did not find any differences between monolinguals and bilinguals in vocabulary comprehension or vocabulary production for infants (13 months). A recent study showed that

multilingualism does not affect infants' gesture repertoires either (Germain et al., 2022). Other studies suggest that multilingual toddlers do not have smaller vocabularies than their monolingual peers when they receive at least 60% exposure to the assessed language (Cattani et al., 2014). In our study, we included multilingual children that have lived in the Netherlands since birth to examine whether they are negatively affected when examining only one of their languages, namely Dutch, using the N_{YOUth}-CDIs and PPVT-III-NL.

3.6.5. Research aim

We aimed to examine whether key predictors that explain variation in children's early vocabularies are age-specific and task-specific in this large, longitudinal sample of Dutch children. A limited number of studies have examined the effects of key predictors of variation in children's vocabularies —maternal education, children's gender, gestational age and birthweight, and multilingualism — within large, longitudinal samples, while there are uncertainties regarding their effects on different vocabulary outcomes across development. By examining the effects on multiple vocabulary measures in a large sample from infancy to toddlerhood, we analyse whether the effects of well-known predictors are age-specific and task-specific while keeping the characteristics of the sample constant.

3.7. Methods

3.7.1. Sample

The same 444 participants from the YOUth cohort study described in Part 1 were included in this study, along with their mothers. In total, 426 mothers filled out the demographics questionnaire including questions on the caregivers' education. All caregivers provided us with their child's due date and birth date which we used to determine the children's gestational duration. Of this sample, 399 caregivers also provided us with their child's birthweight in grams. Lastly, 369 caregivers filled out the questionnaire including languages spoken at home. The summary of sample characteristics is shown in Table 3.3.

In this sample, at least 29 children were not growing up as monolingual Dutch speakers. We considered a child monolingual when only Dutch was spoken at home. Given the small number of multilingual children, we did not differentiate the group further based on the children's estimated time of exposure to Dutch.

Table 3.3. Sample characteristics including the mean (and standard deviation) for continuous variables or frequency counts (and percentage of sample) for categorical variables.

	N	*Mean (SD) or n (% of sample)*
Age in weeks		
Wave 1	338	46.11 (3.79)
Wave 2	444	175.20 (41.48)
Male	444	214 (48%)
Highest maternal education	426	
Primary school		1 (<1%)
High school		16 (4%)
Vocational education		60 (14%)
Higher education		143 (34%)
University education		215 (50%)
Gestational duration in days	444	278 (12)
Birthweight in grams	399	3514 (476)
Multilingual	369	29 (8%)

3.7.2. Materials and procedure

We collected the previously described characteristics of the sample via questionnaires. These included questionnaires on the mother's demographics (e.g., educational background), which we collected when the mother was 20 weeks pregnant; the child's birth (e.g., due date, birth date, and birthweight), which we collected shortly after the child's birth; and the languages spoken at home (including questions about the

caregivers' native language(s) and the language(s) spoken at home), which we collected during Wave 1 concurrently with the NYOUth-CDI 1.

3.7.3. Coding and analysis

We coded the highest educational degree obtained on a nine-point scale ranging from 1 = no education to 9 = university degree. We calculated gestational duration in days using the discrepancy between children's due dates and birth dates and adding or subtracting this from 280 days (i.e., full-term gestation). Caregivers reported their children's birthweight in grams. Lastly, we determined whether a child was growing up multilingual (i.e., at least one caregiver does not only speak Dutch at home).

We fitted robust generalised linear models using the package *robustbase* version 0.95-0 (Maechler et al., 2022) following Frank et al. (2021). We used vocabulary comprehension, vocabulary production, and gestures measured by the NYOUth-CDI 1, production measured by the NYOUth-CDI 2, and vocabulary comprehension measured by the PPVT-III-NL as continuous outcome measures. We added children's ages in weeks, gender (female or male), gestational, birthweight, maternal education, and language status (monolingual or multilingual) as predictors to the models. For categorical predictors, we used dummy coding with the categories containing the largest number of observations (gender: female; language status: monolingual) as reference levels. We centred and scaled children's age, gestational duration, birthweight, and maternal education. We modelled raw scores instead of normed scores or percentiles. By adding age in weeks as a predictor to the models, all other predictors are independent of the effects of age.

3.8. Results

During Wave 1, vocabulary comprehension, vocabulary production, and gestures were measured with the NYOUth-CDI 1. The results of the robust regression models for vocabulary outcomes at Wave 1 are presented in Table 3.4. During Wave 2, vocabulary production was measured with the NYOUth-CDI 2 and comprehension was measured with the PPVT-III-NL. The results of the robust regression models for

vocabulary outcomes at Wave 2 are presented in Table 3.5. When examining the effects on all vocabulary outcomes of infants and toddlers, we find one consistent predictor: age in weeks has a positive effect on all collected outcomes. We expected a robust age-related effect as children's vocabularies grow fast during the first years of development. Figures 3.2 show the effects of children's age and gender on the different N_{YOUth}-CDI 1 scales.

Several other relevant findings emerge. Except for age in weeks, all other predictors show inconsistent patterns across the different measurement waves and vocabulary outcomes. We found a negative effect of maternal education on caregiver-reported vocabulary comprehension and vocabulary production for infants at Wave 1. The negative effect of maternal education has shifted to a positive effect on the lab-administered PPVT-III-NL task during Wave 2, but not on caregiver-reported vocabulary production during this wave.

For children's gender, we only found an advantage for girls on gestures in the N_{YOUth} -CDI 1. During Wave 2, we found an advantage for girls on N_{YOUth}-CDI 2 production, but not on the PPVT-III-NL at this age. Figure 3.2A shows the effect of children's age in weeks and gender on N_{YOUth}-CDI 2 and Figure 3.2B on the PPVT-III-NL at Wave 2. Although both show a similar increase with age, there is a ceiling effect for production measured using the N_{YOUth}-CDI 2.

Table 3.4. Robust regression coefficients with 95% confidence intervals (CIs) for vocabulary outcomes at Wave 1.

	Outcome variable (95% CI)		
	NYOUth-CDI 1 Comprehension	NYOUth-CDI 1 Production	NYOUth-CDI 1 Gestures
(Intercept)	38.66***	1.79***	19.01***
	(35.42, 41.89)	(1.15, 2.44)	(17.98, 20.04)
Age in weeks	8.16***	0.71**	3.60***
	(5.81, 10.51)	(0.26, 1.16)	(2.91, 4.29)
Maternal education	-4.32**	-0.37*	-0.26
	(-7.21, -1.43)	(-0.67, -0.08)	(-0.99, 0.46)
Gender (male)	-4.42	-0.30	-2.60***
	(-9.33, 0.49)	(-0.81, 0.21)	(-4.03, -1.18)
Gestational duration	-0.12	0.04	0.79
	(-3.09, 2.84)	(-0.32, 0.41)	(-0.09, 1.67)
Birthweight	0.53	-0.06	0.25
	(-1.81, 2.86)	(-0.35, 0.24)	(-0.47, 0.98)
Language status (multilingual)	-2.30	0.36	-0.10
	(-10.90, 6.30)	(-0.76, 1.47)	(-2.55, 2.35)
Observations	312	312	311
R^2	0.19	0.14	0.32
Adjusted R^2	0.17	0.12	0.31
Residual Std. Error	19.47 (df = 305)	1.68 (df = 305)	5.54 (df = 304)

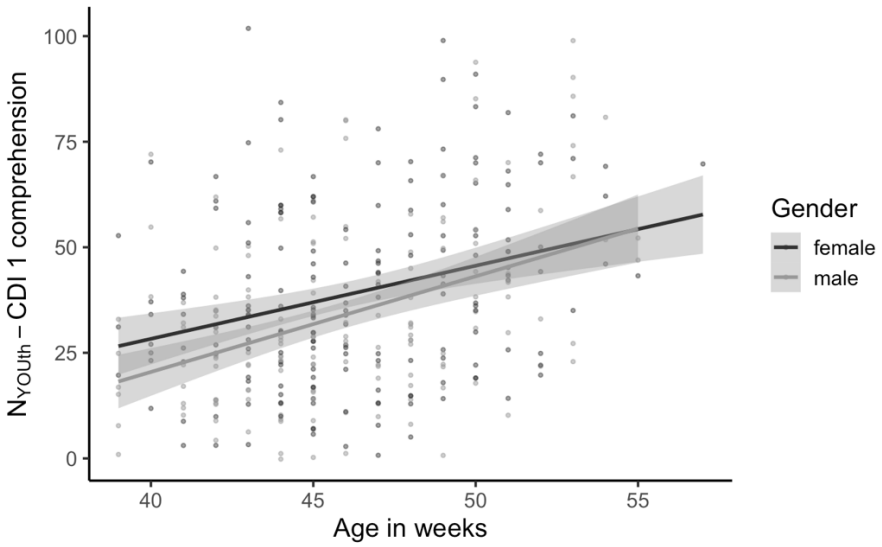
Note: * $p < .05$; ** $p < .01$; ***$p < .001$

Table 3.5. Robust regression coefficients with 95% confidence intervals (CIs) for vocabulary outcomes at Wave 2.

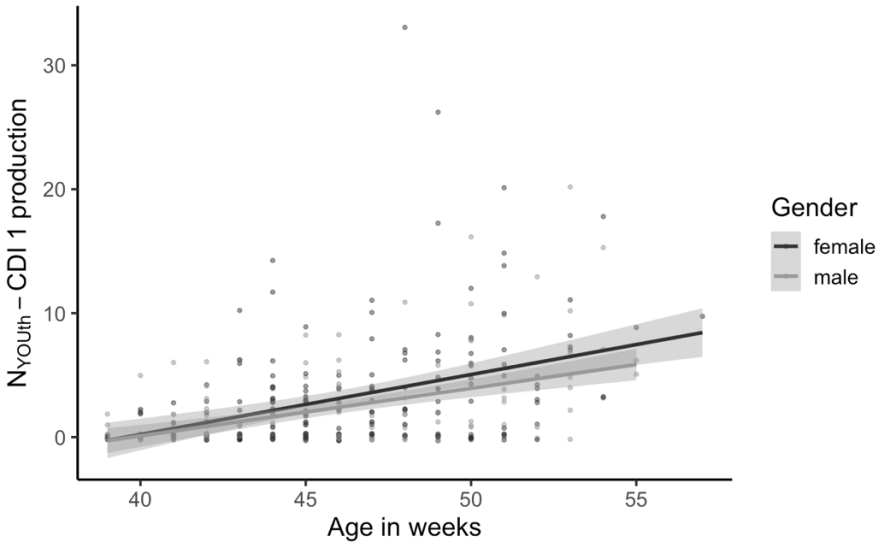
	Outcome variable (95% CI)	
	N_{YOUth}-CDI 2 Production	PPVT-III-NL Comprehension
(Intercept)	183.41***	54.30***
	(179.96, 186.86)	(52.87, 55.74)
Age in weeks	14.85***	14.64***
	(11.55, 18.14)	(13.36, 15.92)
Maternal education	0.18	1.70**
	(-3.53, 3.89)	(0.42, 2.98)
Gender (male)	-5.65*	-1.06
	(-10.59, -0.70)	(-3.28, 1.15)
Gestational duration	-1.42	-0.33
	(-4.65, 1.80)	(-1.87, 1.21)
Birthweight	1.54	1.30
	(-1.60, 4.67)	(-0.50, 3.11)
Language status (multilingual)	-8.11	-5.23*
	(-20.03, 3.80)	(-10.16, -0.30)
Observations	264	325
R^2	0.39	0.68
Adjusted R^2	0.38	0.68
Residual Std. Error	16.95	9.58
	(df = 257)	(df = 318)

Figure 3.2. Effects of children's age and gender on vocabulary comprehension (A), production (B), and gestures (C) measured with the NYOUth-CDI 1 including linear regression lines with 95% confidence intervals.

A. Comprehension (NYOUth-CDI 1)



B. Production (NYOUth-CDI 1)



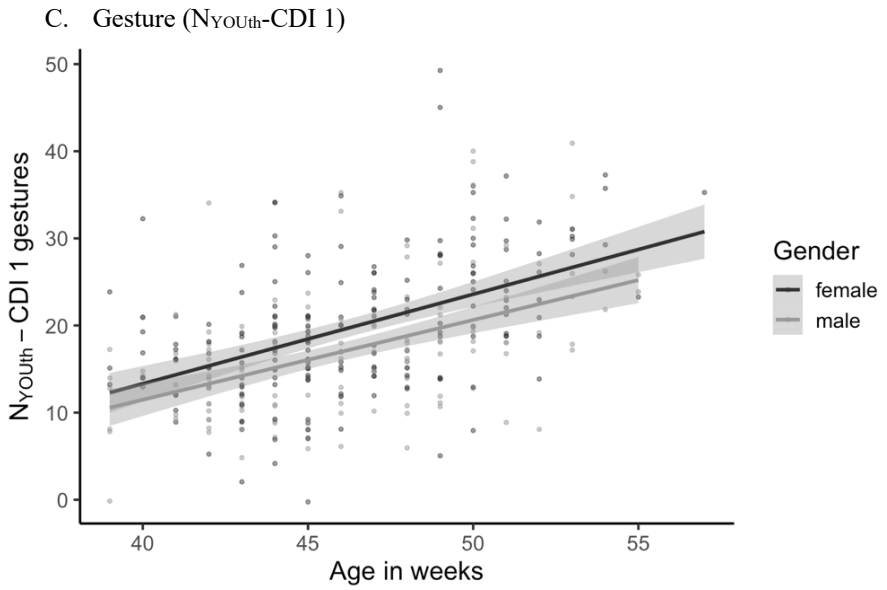
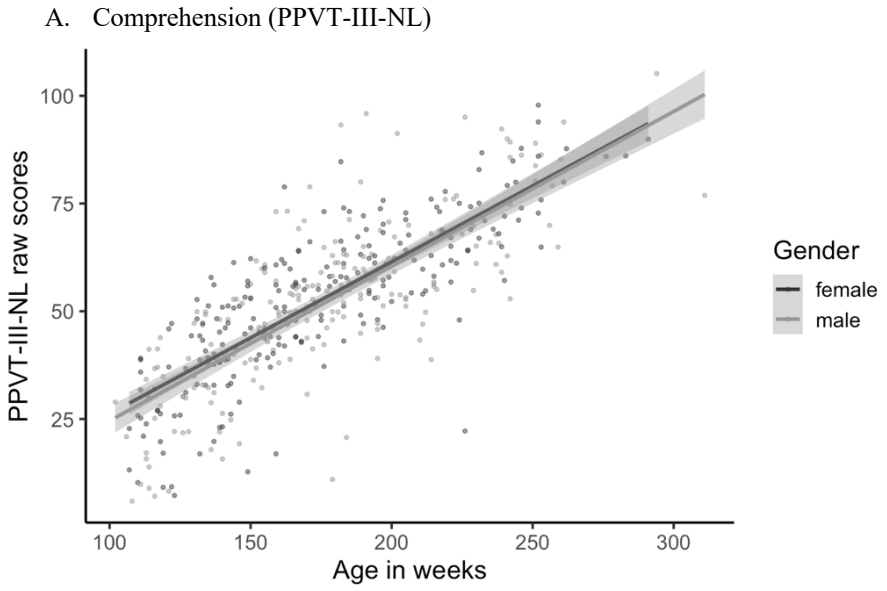
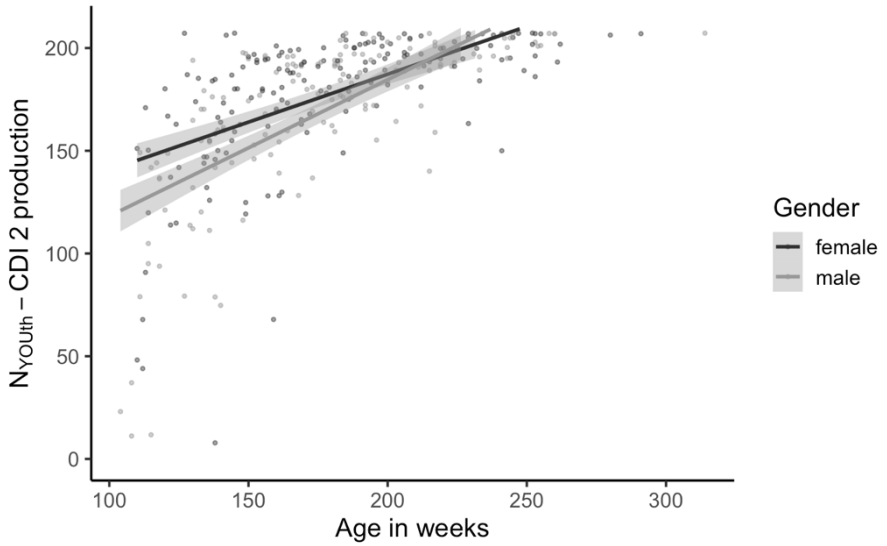


Figure 3.3. Effects of children's age and gender on vocabulary comprehension measured with the PPVT-III-NL (A) and production measured with the NYOUth-CDI 2 (B) including linear regression lines with 95% confidence intervals.



B. Production (N_{YOUth}-CDI 2)



We did not find any effects of gestational age or birthweight on any of the vocabulary outcomes during Wave 1 or Wave 2. Lastly, we only found a negative effect of multilingualism on the PPVT-III-NL, but not on any of the N_{YOUth}-CDIs. These results are discussed below.

3.9. Discussion

We aimed to examine whether key predictors that explain variation in children's early vocabularies — maternal education, children's gender, gestational age and birthweight, and multilingualism — are age-specific and task-specific in this large, longitudinal sample of Dutch children. There is large individual variability across children's vocabulary outcomes. One factor that consistently explains some variation in children's early vocabularies is their age in weeks. This confirms that children's vocabularies progressively develop over time. Below, we address all other factors one by one.

3.9.1. Effect of maternal education shifts over time

We examined the effects of maternal education (as a proxy for SES) on vocabulary outcomes. First, we found negative effects of maternal education on vocabulary production and vocabulary comprehension measured by the NYOUth-CDI 1. This is in line with previous studies that have also reported negative effects of SES on CDIs for infants – usually for vocabulary comprehension and to a lesser extent for word production (Fenson et al., 1994; Reese & Read, 2000; Feldman et al., 2000). This is likely due to caregiver reporting bias. The latter interpretation is strengthened by the finding that there is no effect of maternal education on NYOUth-CDI 1 gestures or NYOUth-CDI 2 production. Gestures and spoken words may be more easily observable and do not require as much interpretation, making them less susceptible to reporting biases. Unlike gestures, word production still requires a small amount of interpretation because caregivers are instructed to also check *speaks* for vocabulary items when their child produces synonyms or production errors. In addition, caregivers may over-report their child's vocabulary if they think larger vocabularies are desirable. This social stigma is less prominent for children's gesture repertoires which makes them less susceptible to reporting biases. Lastly, we found a positive effect of maternal education on PPVT-III-NL scores at Wave 2. This result is in line with previous studies finding that a higher SES, often measured through maternal education, correlates with larger vocabularies (Huttenlocher et al., 2010; Hoff, 2003). This could suggest that an advantage of maternal education only emerges later in children's development, although an effect on infants' vocabularies could be obscured by caregiver reporting biases or floor effects.

3.9.2. Girls have an advantage over boys

The results show that girls have an advantage over boys on NYOUth-CDI 1 gestures and NYOUth-CDI 2 production. Previous studies have also frequently reported an advantage for girls using CDIs (Frank et al., 2021; Reese & Read, 2000; Eriksson et al., 2012; Feldman et al., 2005). The results of our study suggest that the gender difference could start with a difference in children's gesture repertoires since gestures are known to influence children's later vocabularies (see Brooks & Meltzoff, 2008; Colonnesi et

al., 2010; McGillion et al., 2017; Rowe & Goldin-Meadow, 2009). Recently, Germain et al. (2022) also showed that 14-month-old girls produce more gesture types than boys using caregiver reports. Our results add to this finding by showing that a difference in gestures between boys and girls is already present before their first birthday. Our findings are also in line with the hypothesis that gender differences are more prevalent in vocabulary production than vocabulary comprehension (e.g., Frank et al., 2021; Feldman et al., 2005). This could explain the absence of a significant gender effect on the PPVT-III-NL. Another possible explanation for this is that the gender effect on N_{YOUth}-CDIs is the result of a reporting bias. Caregivers could expect that girls are more verbal than boys, influencing how they fill out the vocabulary checklist. However, we suspect that this is unlikely since we found a significant gender effect on word production during Wave 2. Caregiver reports on word production (rather than comprehension) and toddlers (rather than infants) are less susceptible to reporting biases. Frank et al. (2021) also showed that cross-linguistically, the advantage for girls is more prominent in caregiver reports of word production than word comprehension. This suggests that girls truly have an advantage over boys — at least in their expressive vocabularies.

3.9.3. No effect of gestational duration or birthweight

We did not find any effects of gestational duration or birthweight on children's vocabularies in this non-clinical sample. This does not support earlier findings that preterm infants are at risk of having smaller vocabularies later in life than full-term infants (e.g., Foster-Cohen et al., 2007; Guarini et al., 2009; Sansavini et al., 2011). Nevertheless, some studies suggest that only extremely preterm children (under 28 weeks) and/or children of very low birthweight (under 1500 g) have language delays (Barre et al., 2011; Kern & Gayraud, 2007). None of the children included in our rather homogeneous sample fall under those criteria. Therefore, it is possible that we did not find any differences because gestational duration and birthweight predominantly affect the more extreme cases. Future studies should examine the effects of gestational duration and birthweight in longitudinal samples that include very to extremely preterm children and/or children of very low birthweights.

3.9.4. Multilinguals know fewer words than monolinguals

We lastly examined the effects of children in the Netherlands growing up with more than one language. The results show that monolingual toddlers have larger receptive vocabularies measured with the PPVT-III-NL, but not larger productive vocabularies measured with the N$_{YOUth}$-CDIs. Given the fact that multilingual toddlers are not exposed to as much Dutch language input as their monolingual peers, and vocabulary development is heavily influenced by the quantity and quality of exposure (Hoff, 2003), we expected multilingual toddlers to have smaller vocabularies when measuring only one of their languages. In the N$_{YOUth}$-CDIs, caregivers were instructed to also check *speaks* on vocabulary items when their child produces a synonym. Arguably, these instructions yielded large variability in how multilingual caregivers filled out the checklists. It is plausible that some multilingual caregivers also accepted translations for vocabulary items which could explain the absence of a negative effect of multilingualism on the N$_{YOUth}$-CDIs. Our sample could also have been too homogeneous because all caregivers who participated in the YOUth cohort study were required to be able to fill out Dutch questionnaires to participate. This resulted in a small number of multilingual children in our sample that may not have been sufficient to detect an effect of multilingualism on caregiver reports, especially given the potential variability in how multilingual caregivers filled out the reports. Lastly, we found no effect of multilingual input on gestures, which is in line with a recent study that did not find an effect of multilingualism on 14-month-old infants' gestures measured with CDIs (Germain et al., 2022). Even though infants' gesture repertoires are an early indicator of their later vocabulary size, they are likely independent of specific language exposure and therefore not affected by multilingual language input.

3.10. General discussion

In Part 1, we examined the concurrent and predictive validity of the N$_{YOUth}$-CDIs used in the YOUth cohort. After establishing their validity, we examined whether an array of well-known environmental predictors of variation in children's vocabularies were time-specific and task-specific in Part 2. When combining the results of both parts, several relevant findings emerge. First, we propose that the gesture scale provides a

valid measure of infants' vocabularies. Second, the results show that predictors vary across children's ages and vocabulary tasks. This has methodological implications for future studies and provides new insights into the validity of the different vocabulary measurements.

The gesture scale provides a valid measure of infants' vocabularies. In Part 1, we found that the gesture scale was the only infant measure which significantly correlated with later PPVT-III-NL scores. This is particularly strong evidence of the predictive validity of the gesture scale. In Part 2, we found negative effects of maternal education on infants' word production and word comprehension, but not on gestures. This is in line with the findings by Rowland et al. (2022) who found that the reverse SES effect for infants was far less prevalent in the gesture scale across ten cross-linguistic CDI datasets. We also found the expected advantage for girls in the gesture scale, but not in word production or word comprehension during infancy. The gesture scale could be the only scale that shows enough variability across infants, resulting in sufficient variation to detect the gender effect. The gender effect could also manifest itself in gestures first, subsequently influencing children's later vocabularies (e.g., Colonnesi et al., 2010). We lastly found that the gesture scale is not affected by multilingual language input, making this a useful report to collect for multilingual infants as well. Therefore, the gesture scale appears to be particularly valid for capturing individual variation in infants' vocabulary skills. Yet, the gesture scale has an age-range limitation. The predictive power of gestures likely diminishes during the second year of life as most children will have acquired most gestures on the checklist, resulting in a ceiling effect. Nevertheless, the results of our study show that at least for infants around 10 months, both early and late gestures included in the CDIs have more predictive value for children's later vocabularies than word production or word comprehension.

The second finding that emerged is that the effects of well-known predictors of variation in children's vocabularies varied across children's ages and vocabulary outcomes. Apart from a consistent positive effect of children's ages on all outcomes,

we found that none of the other predictors remained constant across the different vocabulary outcomes measured in this study. During infancy, there is a negative effect of maternal education on caregiver-reported word production and word comprehension which has disappeared for word production in toddlerhood. The bias is likely more prominent in infants since caregiver reports of infants' skills require more interpretation, and they are likely more influenced by caregivers trying to meet certain expectations surrounding their infants' development. For infants, we found an advantage for girls on the gesture scale. For toddlers, we found an advantage for girls for expressive (measured with the NYOUth-CDI 2) but not receptive language (measured with the PPVT-III-NL). Although this could indicate that the caregiver-reported measure is affected by a reporting bias, multiple studies using different measurement instruments showed that the advantage for girls is more prominent in children's expressive language (e.g., Bornstein et al., 1998; Frank et al., 2021; Feldman et al., 2005; Qi et al., 2003). Therefore, we suggest that the advantage found for girls on the NYOUth-CDI 2 supports its validity. Despite the ceiling effect, there is still enough variability to capture the effect. Lastly, we found that multilingual language input negatively affects the lab-administered PPVT-III-NL but none of the NYOUth-CDIs. Based on previous research, we would expect multilingual toddlers to have smaller vocabularies than their monolingual peers when assessing only one of their languages (Blom et al., 2020; De Houwer et al., 2014; Hoff et al., 2012). Therefore, we suggest that this negatively affects the validity of NYOUth-CDIs for multilingual children when the goal is to capture their vocabulary size in one language.

3.10.1. Limitations and future studies

Although we found some effects of maternal education in the expected directions based on previous studies using socio-demographically diverse samples (Feldman et al., 2000), our sample is rather homogeneous and overrepresents highly educated mothers. A lack of diversity makes SES differences less apparent. The results of our study suggest that caregiver reports of infants' vocabulary comprehension and vocabulary production, but not gestures, are negatively affected by maternal education. According to previous studies, infants of lower SES may be using fewer

gestures during caregiver-child interactions (Rowe & Goldin-Meadow, 2009). Although we did not examine gesture rates, we did not find an effect of maternal education on infants' gesture repertories. Future longitudinal cohort studies with more diverse samples should re-evaluate whether gesture repertoires show any SES differences and how these may affect the predictive value of gestures in diverse samples.

3.11. Conclusions

We conclude that the N_{YOUth}-CDIs are a cheap, fast, reliable, and valid method to capture variability in infants' and toddlers' vocabularies. However, the predictive validity for infants around 10 months of age is limited. For this age group, we would recommend including gestures when administering caregiver reports. The results of our study provide evidence that caregiver reports of gestures are a relevant measure of infants' early vocabularies. Then, the results of our longitudinal study including over 300 Dutch children additionally suggest that the effects of key predictors on children's vocabularies are dependent on children's ages and vocabulary outcomes. One advantage of cohort studies is to gain better insights into which predictors have temporary or weak effects on development. Given our results, we would recommend sampling diverse samples of children and use more than one vocabulary measure when examining predictors of individual variation to gain a more comprehensive understanding of any effects on vocabulary. Well-known predictors can differentially affect children's gestures during infancy and their expressive and comprehensive vocabularies across development. We also found that effects can shift over time, at least from infancy to toddlerhood. This corroborates that we should examine large, longitudinal samples cross-linguistically to determine the generalisability and stability of key predictors on vocabulary development.

Acknowledgements and data availability

We are grateful to all families who participate in the YOUth study. YOUth is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant

number 024.001.003). A complete listing of the study investigators and study management can be found at https://www.uu.nl/en/research/youth-cohort-study/about-us/who-isinvolved. YOUth investigators and management designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the YOUth study investigators or YOUth management. YOUth is a longitudinal study that aims to produce and safely store FAIR and high-quality data. The data can be accessed for both use and verification purposes upon request (see https://www.uu.nl/en/research/youth-cohort-study/data-access). The *R* script and other materials can be found online: https://osf.io/vj72c/

References

Barre, N., Morgan, A., Doyle, L. W., & Anderson, P. J. (2011). Language abilities in children who were very preterm and/or very low birth weight: A meta-analysis. *The Journal of Pediatrics*, *158*(5), 766-774.e1. https://doi.org/10.1016/j.jpeds.2010.10.032

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In *The handbook of child language* (pp. 95–151). Blackwell.

Bavin, E. L., Prior, M., Reilly, S., Bretherton, L., Williams, J., Eadie, P., Barrett, Y., & Ukoumunne, O. C. (2008). The Early Language in Victoria Study: Predicting vocabulary at age one and two years from gesture and object use. *Journal of Child Language*, *35*(3), 687–701. https://doi.org/10.1017/S0305000908008726

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*(7080), 572. https://doi.org/10.1136/bmj.314.7080.572

Blom, E., Boerma, T., Bosma, E., Cornips, L., van den Heuij, K., & Timmermeister, M. (2020). Cross-language distance influences receptive vocabulary outcomes of bilingual children. *First Language*, *40*(2), 151–171. https://doi.org/10.1177/0142723719892794

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*(2), 367–393. https://doi.org/10.1017/s0305000998003456

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/s030500090700829x

Cadime, I., Silva, C., Santos, S., Ribeiro, I., & Viana, F. L. (2017). The interrelatedness between infants' communicative gestures and lexicon size: A longitudinal study. *Infant Behavior and Development*, *48*, 88–97. https://doi.org/10.1016/j.infbeh.2017.05.005

Cattani, A., Abbot-Smith, K., Farag, R., Krott, A., Arreckx, F., Dennis, I., & Floccia, C. (2014). How much exposure to English is necessary for a bilingual toddler to perform like a monolingual peer in language tests? *International Journal of Language & Communication Disorders*, *49*(6), 649–671. https://doi.org/10.1111/1460-6984.12082

Colonnesi, C., Stams, G. J. J. M., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366. https://doi.org/10.1016/j.dr.2010.10.001

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345. https://doi.org/10.1111/cdev.12193

De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual–monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied Psycholinguistics*, *35*(6), 1189–1211. https://doi.org/10.1017/S0142716412000744

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-III*. American Guidance Service.

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M., & Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology*, *30*(2), 326–343. https://doi.org/10.1111/j.2044-835X.2011.02042.x

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 Years. *Child Development*, *76*(4), 856–868. https://doi.org/10.1111/j.1467-8624.2005.00882.x

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*(2), 310–322. https://doi.org/10.1111/1467-8624.00146

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. https://doi.org/10.2307/1166093

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur Communicative Development Inventories: User's guide and technical manual* (Second edition). Paul H. Brookes Publishing Co., Inc.

Foster-Cohen, S., Edgin, J. O., Champion, P. R., & Woodward, L. J. (2007). Early delayed language development in very preterm infants: Evidence from the MacArthur-Bates CDI*. *Journal of Child Language*, *34*(3), 655–675. https://doi.org/10.1017/S0305000907008070

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank Project*. MIT Press. https://langcog.github.io/wordbank-book/

Germain, N., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2022). Gesture development in infancy: Effects of gender but not bilingualism. *Infancy*, *27*(4), 663–681. https://doi.org/10.1111/infa.12469

Guarini, A., Sansavini, A., Fabbri, C., Alessandroni, R., Faldella, G., & Karmiloff-Smith, A. (2009). Reconsidering the impact of preterm birth on language outcome. *Early Human Development*, *85*(10), 639–645. https://doi.org/10.1016/j.earlhumdev.2009.08.061

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, *39*(1), 1–27. https://doi.org/10.1017/S0305000910000759

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *Journal of Child Language*, *34*(4), 701–724. https://doi.org/10.1017/S0305000907008124

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*, 236–248. https://doi.org/10.1037/0012-1649.27.2.236

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. https://doi.org/10.1016/j.cogpsych.2010.08.002

Kern, S., & Gayraud, F. (2007). Influence of preterm birth on early lexical and grammatical acquisition. *First Language*, *27*(2), 159–173. https://doi.org/10.1177/0142723706075790

Kidd, E., & Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics*, *6*(1), 319–340. https://doi.org/10.1146/annurev-linguistics-011619-030326

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in infant speech segmentation: Achieving the lexical shift. *Infancy*, *23*(6), 770–794. https://doi.org/10.1111/infa.12256

Kim, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, *22*(6), 665. https://doi.org/10.5351/CSAM.2015.22.6.665

Kuvač Kraljević, J., Cepanec, M., & Šimleša, S. (2014). Gestural development and its relation to a child's early vocabulary. *Infant Behavior and Development*, *37*(2), 192–202. https://doi.org/10.1016/j.infbeh.2014.01.004

Kuvač-Kraljević, J., Blaži, A., Schults, A., Tulviste, T., & Stolt, S. (2021). Influence of internal and external factors on early language skills: A cross-linguistic study. *Infant Behavior and Development*, *63*, 101552. https://doi.org/10.1016/j.infbeh.2021.101552

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2022). *robustbase: Basic Robust Statistics*. http://robustbase.r-forge.r-project.org/

McGillion, M., Herbert, J. S., Pine, J., Vihman, M., dePaolis, R., Keren-Portnoy, T., & Matthews, D. (2017). What paves the way to conventional language? The predictive value of babble, pointing, and socioeconomic status. *Child Development*, *88*(1), 156–166. https://doi.org/10.1111/cdev.12671

O' Toole, C., & Fletcher, P. (2010). Validity of a parent report instrument for Irish-speaking toddlers. *First Language*, *30*(2), 199–217. https://doi.org/10.1177/0142723709359237

Olson, J., & Masur, E. F. (2015). Mothers' labeling responses to infants' gestures predict vocabulary outcomes. *Journal of Child Language*, *42*(6), 1289–1311. https://doi.org/10.1017/S0305000914000828

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., van der Wal, J. J. M., Hulshoff Pol, H. E., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, *46*, 100868. https://doi.org/10.1016/j.dcn.2020.100868

Pan, B. A., Rowe, M. L., Spier, E., & Tamis-Lemonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language*, *31*(3), 587–608. https://doi.org/10.1017/S0305000904006270

Pérez-Pereira, M., & Cruz, R. (2018). A longitudinal study of vocabulary size and composition in low risk preterm children. *First Language*, *38*(1), 72–94. https://doi.org/10.1177/0142723717730484

Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology*, *115*, 101238. https://doi.org/10.1016/j.cogpsych.2019.101238

Qi, C. H., Kaiser, A. P., Milan, S. E., Yzquierdo, Z., & Hancock, T. B. (2003). The performance of low-income, African American children on the Preschool Language Scale—3. *Journal of Speech, Language, and Hearing Research*, *46*(3), 576–590. https://doi.org/10.1044/1092-4388(2003/046)

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur Communicative Development Inventory: Words and Sentences. *Journal of Child Language*, *27*(2), 255–266. https://doi.org/10.1017/S0305000900004098

Reilly, S., Bavin, E. L., Bretherton, L., Conway, L., Eadie, P., Cini, E., Prior, M., Ukoumunne, O. C., & Wake, M. (2009). The Early Language in Victoria Study (ELVS): A prospective, longitudinal study of communication skills and expressive vocabulary development at 8, 12 and 24 months. *International Journal of Speech-Language Pathology*, *11*(5), 344–357. https://doi.org/10.1080/17549500903147560

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*.

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rowland, C., Krajewski, G., Meints, K., Łuniewska, M., Kochańska, M. K., & Alcock, K. (2022). *CDI Demographics*. https://osf.io/hwg4c/

Sansavini, A., Bello, A., Guarini, A., Savini, S., Stefanini, S., & Caselli, M. C. (2010). Early development of gestures, object-related-actions, word comprehension and word production, and their relationships in Italian infants: A longitudinal study. *Gesture*, *10*(1), 52–85. https://doi.org/10.1075/gest.10.1.04san

Sansavini, A., Guarini, A., Savini, S., Broccoli, S., Justice, L., Alessandroni, R., & Faldella, G. (2011). Longitudinal trajectories of gestural and linguistic abilities in very preterm infants in the second year of life. *Neuropsychologia*, *49*(13), 3677–3688. https://doi.org/10.1016/j.neuropsychologia.2011.09.023

Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL*. Hartcourt Assessment BV.

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

Tomasello, M., & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in Child Development*, *59*, 174–179. https://doi.org/10.1111/j.1540-5834.1994.tb00186.x

van Baar, A. L., Ultee, K., Gunning, W. B., & Soepatmi, S. (2006). Developmental course of very preterm children in relation to school outcome. *Journal of Developmental and Physical Disabilities*, *18*(3), 273–293. https://doi.org/10.1007/s10882-006-9016-6

Verhoef, E., Shapland, C. Y., Fisher, S. E., Dale, P. S., & Pourcain, B. S. (2021). The developmental genetic architecture of vocabulary skills during the first three years of life: Capturing emerging associations with later-life reading and cognition. *PLOS Genetics*, *17*(2), e1009144. https://doi.org/10.1371/journal.pgen.1009144

Washington, J. A., & Craig, H. K. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools*, *30*(1), 75–82. https://doi.org/10.1044/0161-1461.3001.75

Wu, Z., & Gros-Louis, J. (2015). Caregivers provide more labeling responses to infants' pointing than to infants' object-directed vocalizations*. *Journal of Child Language*, *42*(3), 538–561. https://doi.org/10.1017/S0305000914000221

Zink, I., & Lejaegere, M. (2002). *N-CDI's: Lijsten voor Communicatieve Ontwikkeling. Aanpassing en hernormering van de MacArthur CDI's van Fenson et al.* Acco.

Zink, I., & Lejaegere, M. (2003). *N-CDI's: Korte vormen, Aanpassing en hernormering van de MacArthur Short Form Vocabulary Checklist van Fenson et al.* Acco.

Chapter 4

Infants' behaviours elicit different verbal, nonverbal, and multimodal responses from caregivers during early play

Abstract

Caregivers use a range of verbal and nonverbal behaviours when responding to their infants. Previous studies have typically focused on the role of the caregiver in providing verbal responses, while communication is inherently multimodal (involving audio and visual information) and bidirectional (exchange of information between infant and caregiver). In this paper, we present a comprehensive study of caregivers' verbal, nonverbal, and multimodal responses to 10-month-old infants' vocalisations and gestures during free play. A new coding scheme was used to annotate 2,036 infant vocalisations and gestures of which 87.1% received a caregiver response. Most caregiver responses were verbal, but 39.7% of all responses were multimodal. We also examined whether different infant behaviours elicited different responses from caregivers. Infant bimodal (i.e., vocal-gestural combination) behaviours elicited high rates of verbal responses and high rates of multimodal responses, while infant gestures elicited high rates of nonverbal responses. We also found that the types of verbal and nonverbal responses differed as a function of infant behaviour. The results indicate that infants influence the rates and types of responses they receive from caregivers. When examining caregiver-child interactions, analysing caregivers' verbal responses alone undermines the multimodal richness and bidirectionality of early communication.

The full chapter has been published as:

van der Klis, A., Adriaans, F., & Kager, R. (2023). Infants' behaviours elicit different verbal, nonverbal, and multimodal responses from caregivers during early play. *Infant Behavior and Development*, *71*, 101828.

4.1. Introduction

During early play sessions, infants may babble while pointing at a doll. Their caregiver may pick up the doll and ask: "Do you want this doll?". The infant starts reaching for the doll, extending both arms while opening and closing their fingers. Their caregiver smiles in understanding and hands over the doll. Such interactions between infants and caregivers help the infant to identify the label "doll" for the object they were interested in. More generally, such interactions teach the infant to communicate effectively by producing sounds and gestures. Previous studies have shown that caregiver responses to infant vocalisations and gestures differ in terms of frequency and contents (e.g., Ger et al., 2018; McGillion et al., 2013; Olson & Masur, 2013; Tamis-LeMonda & Bornstein, 2002; Wu & Gros-Louis, 2014). However, while communication is inherently multimodal (involving audio and visual information) and bidirectional (exchange of information between infant and caregiver), previous studies have mostly focused on verbal responses provided by the caregiver. It is therefore not known to what extent caregivers produce nonverbal or multimodal responses, and to what extent such caregiver responses are elicited by different infant behaviours. Understanding the full extent of early infant-caregiver interactions is crucial for understanding infants' early socio-cognitive development. The current study has taken an important step towards this goal by investigating caregivers' verbal, nonverbal, and multimodal responses to 10-month-old infants' vocalisations and gestures during free play.

4.1.1. Infants learn to communicate

Infants learn to communicate by producing vocalisations and gestures. A study by Donnellan et al. (2019) found that 11-month-olds varied greatly in their vocalisations during a play session at home, with some infants producing two or three vocalisations, and other infants producing more than a hundred vocalisations during the session. This shows that early vocalisations are characterised by large variability. The developmental trajectories of gestures are also cross-linguistically characterised by both variability and stability (Fenson et al., 1994; Frank et al., 2021). The earliest

deictic gestures typically involve giving and showing, later followed by index-finger pointing and requesting, although the ages of onset can vary drastically across children (Frank et al., 2021). These individual differences in the productions of vocalisations and deictic gestures, specifically index-finger pointing, have been positively associated with children's vocabulary outcomes (e.g., Brooks & Meltzoff, 2008; Choi et al., 2021; Colonnesi et al., 2010; Rowe et al., 2022). It remains unclear, however, why individual differences are related to children's later language skills. Some studies have suggested that infant behaviours elicit specific contingent responses from caregivers that facilitate language development (Ger et al., 2018; Olson & Masur, 2015), for example, providing a label for an object that the infant was pointing at. Crucially, infants must first produce vocalisations and gestures to create opportunities for their caregivers to respond. The information flow is in both directions.

4.1.2. Communication is bidirectional

Young infants already expect their caregivers to respond to their prelinguistic vocalisations and gestures. At 5 months of age, infants have learned that vocalisations elicit caregiver responses (Goldstein et al., 2009). By 10-12 months, infants use deictic gestures with the motive to share attention and interest with others (Boundy et al., 2019; Liszkowski et al., 2004). The onset of declarative communication therefore takes place before infants learn how to speak. In the process, infants learn which behaviours are effective at eliciting which types of caregiver responses. In turn, caregivers should be sensitive to differences in infants' communicative behaviours and respond contingently and appropriately. This bidirectional view of communication has recently shifted the focus from studying individual behaviours to examining a shared system in which infants and caregivers both shape the interaction (Chen et al., 2021; Renzi et al., 2017). When examining individual differences in caregivers' language input to their infant, we cannot attribute all variation to the caregivers themselves because infants also influence the input they receive from their caregivers.

In particular, studies have shown that caregiver responses vary as a function of child behaviour (e.g., Choi et al., 2021; Gros-Louis et al., 2006; Olson & Masur, 2013). Olson and Masur (2013) showed that mothers provide more object labels to gestural than non-gestural bids. This shows that infants who produce more gestures tend to elicit more labelling responses from their caregivers. The type of vocalisation or gesture also influences the response. More specifically, infant index-finger pointing gestures have been found to elicit more labelling responses than reaching gestures (Kishimoto et al., 2007; Wu & Gros-Louis, 2015). In a recent study, Choi et al. (2021) showed that caregivers respond more often to their 10-month-old infants' showing + giving gestures than to their pointing gestures. They only examined these gestures and did not distinguish between different types of responses. In addition, mothers were found to use more verbal than nonverbal responses to infant vocalisations, and they responded with more vocal imitations to consonant-vowel (CV) sequences compared to vowel-like sounds (Gros-Louis et al., 2006). Recently, Yurkovic et al. (2021) showed that multimodal behaviours (looks combined with touch) by infants aged 12–48 months elicited higher caregiver response rates than unimodal behaviours and elicited more multimodal (looks combined with touch) responses from caregivers. These studies provide some initial evidence that infant behaviours tend to elicit caregiver responses in the same modality. Symmetry between modalities could suggest high synchrony between children and their caregivers (Leclère et al., 2014), but more research is needed to establish from which age and to what extent caregiver-child dyads match modalities in infant behaviour and caregiver response sequences. The studies so far suggest that certain infant behaviours tend to elicit higher response rates and different response types, but they only examined a few types of infant behaviours and caregiver responses. We currently miss a detailed characterisation of caregivers' verbal, nonverbal, and multimodal behaviours in response to infants' vocalisations, gestures, and bimodal behaviours.

4.1.3. Relevance of verbal and nonverbal responses

Caregivers individually differ in their verbal responsiveness which has been found to positively relate to children's vocabulary development (e.g., Donnellan et al., 2019;

McGillion et al., 2013; Olson & Masur, 2015; Wu & Gros-Louis, 2014). Variation in caregiver responsiveness is rooted in a variety of factors. One factor is socio-economic status (SES). Mothers of a higher SES have been found to produce more speech and verbally respond more often than mothers from lower SES backgrounds (e.g., Hart & Risley, 1995; McGillion et al., 2017; Vanormelingen & Gillis, 2016). It has been suggested that this is the main reason why children from lower SES backgrounds tend to have smaller vocabularies (Huttenlocher et al., 2010), although it is also possible that some infants produce fewer behaviours that elicit verbal responses, giving fewer opportunities for their caregivers to provide contingent language input. A contingent verbal response, such as labelling an object that the infant is pointing at, creates a temporal and semantic contingency that allows the infant to match the phonological form of a word with its meaning. Studies examining individual differences in caregiver responsiveness have therefore largely focused on differences in caregivers' verbal responses.

However, there is evidence that nonverbal responses play a facilitative role as well. Caregivers' responsiveness measured both verbally *and* nonverbally positively relates to their infants' socio-cognitive skills, including language development (see Bornstein & Tamis-LeMonda, 1989). Studies have found that specific caregivers' nonverbal behaviours, such as handing over a toy, pointing, or smiling, predict vocabulary outcomes and social skills (Pearson et al., 2011; Ruddy & Bornstein, 1982). In addition, nonverbal behaviours regularly co-occur with speech. Children appear to rely on visual information when speech is novel (e.g., a label for an unfamiliar object) or unclear (e.g., in the case of referential ambiguity). Studies have found that children use gaze direction, body orientation, and index-finger pointing as cues to learn the reference of novel words from both humans and robots (Baldwin et al., 1996; Grassmann & Tomasello, 2010; Kory Westlund et al., 2017; Verhagen et al., 2019). Recently, Chen et al. (2021) showed that caregivers touched objects more often while naming them when the object was unfamiliar to the child. Overall, approximately 40% of all caregivers' utterances are accompanied by at least one visual cue (Ger et al., 2018; Vigliocco et al., 2019). These studies suggest that caregivers tend to use many

different types of nonverbal cues when providing children with novel speech. Studies have not yet addressed individual differences across caregivers in nonverbal responsiveness, except for one study which found that mothers are more likely to respond verbally, while fathers are equally likely to produce verbal or nonverbal responses (Flippin & Watson, 2011). Existing studies have not yet identified which types of nonverbal and multimodal responses occur during early caregiver-child interactions, nor whether infants also affect their caregivers' nonverbal and multimodal responsiveness.

4.1.4. Current study

Previous studies have documented caregivers' verbal responses in much detail, while different strands of research have shown that nonverbal and multimodal behaviours occur frequently in caregivers' communication with infants. The first aim of this study therefore was to examine which infant vocalisations, gestures, and bimodal behaviours and which caregiver verbal, nonverbal, and multimodal responses occur during early play. For this analysis, we developed a new coding scheme that includes various types of caregivers' nonverbal behaviours, such as gestures, facial expressions, and other non-gestural bodily behaviours, including body orientation. By applying this coding scheme to a large sample of caregiver-child dyads, our study obtained new insights into the richness and variability of early interactions. In addition, while most research focused on the role of the caregiver in providing contingent responses, some studies suggest that infants play a role in eliciting specific caregiver response rates and types. The second aim of this study was to assess whether infants' vocalisations, gestures, and bimodal behaviours elicited different verbal, nonverbal, and multimodal responses. We examined this through statistical analyses of co-occurring infant behaviours and caregiver responses. These analyses informed us to what extent infants affect their caregivers' responsiveness during early play – thereby shaping their own language experience.

4.2. Method

4.2.1. Participants

The data for this study are derived from YOUth, an ongoing longitudinal cohort study that is part of Utrecht University and University Medical Center Utrecht (see Onland-Moret et al., 2020). YOUth has repeated measurements at regular intervals ("waves"). The current study uses measurements obtained at the age of 9–11 months. From the original sample, we excluded 5 dyads due to technical issues during the recordings resulting in unclear or distorted audio/video, and 1 dyad was excluded because the child was vocalising non-stop throughout the entire video. The final sample consisted of 117 infants (66 females) around 9–11 months of age (M = 10.5 months, SD = 0.9) and their caregivers (92 mothers; 25 fathers). These dyads were selected because they spoke Dutch at home and completed the caregiver-child interaction task. The YOUth cohort study is carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki), and all caregivers have signed informed consent. The study was approved by the medical ethical committee of the University Medical Center Utrecht (application number 14-7-221). Children received a picture book after participating.

We collected information regarding the caregivers' languages spoken at home and their education level via questionnaires. All caregivers included in this study speak only Dutch at home. Most caregivers in our sample were willing to share information regarding their education level (94.9%). We rated their highest level of education on a 5-point scale. Most caregivers (81.1%) completed university or college education and another 15.3% completed senior secondary vocational education. Only a small percentage of caregivers (3.6%) did not continue their education after secondary school. We also collected each partner's educational level when applicable. We found that caregivers' educational levels were moderately correlated with their partners' (ρ = .33, p <.01). Thus, most caregivers in this sample are highly educated.

4.2.2. Procedure

During the caregiver-child interaction task, the infant and their caregiver were asked to sit next to each other within touching distance on a playing rug in a sparsely furnished room. Before the task, the research assistant placed various toys on the rug. The positioning of all items was the same for each participant. The toys were directly in front of both the caregiver and the infant on the rug. The research assistant read a specific set of instructions to ensure each participant received the same instructions. Participants were filmed from four camera angles placed around the rug. Three Dome cameras could be controlled (i.e., moved and zoomed in and out) by the research assistant. One camera captured both the infant and their caregiver from the side, one camera focused on the infant's face, and the other camera focused on the caregiver's face. There was also one fixed camera providing an overview of the entire scene. To capture sound, a fixed, standing (Sennheiser ME64/K6P condenser) microphone is positioned next to the rug. After reading out the set of instructions, the research assistant would take place behind a screen so the caregiver and child could not see them while being filmed. Other caregivers or siblings were not allowed in the lab to minimise distractions.

The caregiver and their infant were filmed for a total of fifteen minutes, subdivided into five different tasks. The session started with three minutes of free play with a standard set of toys and ended with three minutes of free play before the toys were cleaned up. The toys were a baby doll with a milk bottle, a green toy car, a Bumba pop-up toy, and a sun-shaped rattle, in addition to a shape sorter and a picture book that were only available in the last three minutes of free play. The caregivers were asked beforehand to carry the infant back to the rug in case they crawled away. Furthermore, after every three-minute episode, the research assistant gave clear instructions for the next task, e.g.: "It is now time for free play.". After reading the instructions, the research assistant started a stopwatch to film for three minutes. For the present study, we analysed the first and last sessions of free play, analysing six minutes of free play per dyad in total.

4.2.3. Coding scheme

In this section, we present a new coding scheme for annotating infants' vocalisations, gestures, and bimodal behaviours and caregivers' responses. This is the first coding scheme to include various types of verbal, nonverbal, and multimodal responses.

Coding infant behaviours

All videos were coded in ELAN 6.0 (Sloetjes & Wittenburg, 2008). Based on previous studies, an infant vocalisation was any sound produced by the infant except vegetative sounds (e.g., hiccoughs or coughs) or distress sounds (e.g., crying or fussing). A vocalisation was coded as CV vocalisation when at least one syllable contained a consonant-vowel sequence ("baba", "ma" etc.), excluding glides ("ja") and glottals ("ha") (following e.g., Donnellan et al., 2019). All other types of vocalisations were coded as non-CV vocalisation. The categories of infant gestures were 1) index-finger pointing, 2) whole-hand pointing, 3) other rudimentary forms of pointing (e.g., using a fist), 4) showing, 5) giving, 6) reaching, 7) requesting, and 8) other conventional gestures, such as waving or nodding. The selection of these gestures was partially based on previous studies examining infant gestures and caregiver responses (e.g., Donnellan et al., 2019; McGillion et al., 2013; Olson & Masur, 2015; Wu & Gros-Louis, 2014) and the gestures included in the widely used checklist to measure infants' vocabularies: The MacArthur-Bates Communicative Development Inventory (CDI) - Words and Gestures (Fenson et al., 2007). In ELAN, the end of each gesture was marked at the frame where the retraction of the arm began. Infant gestures and vocalisations were coded independently as they could occur simultaneously. This automatically revealed which infant behaviours were bimodal (i.e., gesture-vocal combinations). There had to be at least a partial overlap between the vocalisation and gesture. Otherwise, they were annotated as two separate infant behaviours. For the detailed coding scheme and definitions for each behaviour, see Appendix A.

Coding caregiver responses

After the offset of the infant gesture or vocalisation, a period of exactly two seconds was measured for the caregiver response (following e.g., McGillion et al., 2013; Wu

& Gros-Louis, 2014). The onset of the caregiver response had to occur either during the infant's behaviour or within this two-second time frame. If not, the response was not considered temporally contingent and not analysed in this study. First, we annotated a binary measure indicating whether there was a caregiver response or not. This could be any type of response in the coding scheme (i.e., any verbal or nonverbal behaviour). Then in detail, we annotated which types of caregiver responses occurred. This could be any verbal, gestural, facial, and/or bodily response. These four categories were not mutually exclusive: more than one type of behaviour could occur at the same time. *Verbal* responses were coded as either 1) semantically contingent (i.e., a follow-in response), 2) onomatopoeias or sound effects, 3) infant imitations, or 4) any other type of verbal response that was not semantically related (i.e., non-contingent), such as an affirmation. This selection was based on previous studies (e.g., Donnellan et al., 2019; Motamedi et al., 2021; Tamis-LeMonda & Bornstein, 2002). A response was semantically contingent if its semantic content was related to the attentional state of the infant (following Donnellan et al., 2019; McGillion et al., 2017). We assumed the object or activity was the infant's focus of attention when the infant was vocalising while either holding the object or playing with the object, performing the activity, looking at the object, and/or gesturing towards the object. Otherwise, if the verbal response was not onomatopoeic, a sound effect, or a vocal imitation, the verbal response was coded as non-contingent.

For nonverbal responses, we included different types of gestures, facial expressions, and other (non-gestural and non-facial) bodily behaviours. *Gestural* responses included various types of manual and non-manual gestures. They were further subdivided into 1) pointing, 2) passing, 3) showing (i.e., without manipulating the object), 4) accepting, 5) a representational gesture, 6) object manipulations (see Murgiano et al., 2021), or 7) any other conventional gesture (e.g., nodding or waving). We also included representational gestures (i.e., showing the size, shape, or how an object works without the object in hand) and object manipulations (i.e., when the caregiver interacts physically with an object to play with it or communicate about it) based on the ECOLANG project (Vigliocco et al., in prep). For *facial* responses, we

distinguished between 1) smiling (including laughter), 2) surprise, and 3) other facial expressions different from neutral. We included smiling because it is typical of infant-directed speech (see Benders, 2013). We included surprise because we may expect this to occur frequently during early play, as "mock" surprise in the context of, for example, playing peek-a-boo. The latter (other) category was added to ensure that the coding process was exhaustive. A facial expression only counted as a response if the caregiver was not already showing the expression before the start of the infant's behaviour. This was to ensure that the annotated facial expressions were truly responses to infants' behaviours. Finally, all *bodily* responses were subdivided into 1) leaning closer to the infant, 2) turning to the infant, 3) turning to the toy, and 4) any affective behaviour (e.g., hugging or touching the infant). Body orientations were included in the coding scheme because they may serve as referential cues when hearing novel speech (e.g., Kory Westlund et al., 2017). For the full coding scheme including definitions, see Appendix A.

4.2.4. Training, improving, and reliabilities

We had a three-step process to complete data annotation: training the research assistant, improving the coding scheme after a pilot, and checking inter-rater reliabilities. The first author wrote coding instructions and the initial version of the coding scheme. The first author and research assistant annotated the same three randomly selected videos. They verbally went over all annotations to discuss any differences and uncertainties. After that, the assistant and first author both separately annotated the videos again including an additional seven randomly selected videos. To assess inter-annotator reliabilities, we report chance-corrected modified Cohen's kappa (κ) using the built-in calculator in ELAN 6.0 which is based on the EasyDIAg toolbox (Holle & Rein, 2015). The modified kappa considers both the categorisation of behaviours and the temporal overlap of annotations (i.e., segmentation). While this is a good measure of reliability, kappa values are affected by the large number of coding categories and the infrequent occurrence of some codes. When the marginal distributions are not uniform, the maximum value of kappa cannot reach 1.0 (von Eye & von Eye, 2008). Therefore, we also report maximum kappa, which aids

interpretation of the reported kappa values, and raw agreement, representing the number of agreements on cases divided by the total number of cases, which is a more intuitive measure.

For the first ten videos of the pilot, we found high agreement on categorising different behaviours (κ = .87; κ_{max} = .96; raw = .91). When also including unmatched annotations, the overall agreement dropped to a level that is below satisfaction (κ = .35; κ_{max} = 91; raw = .46). This suggests that there was high agreement on the behaviours that were annotated by both annotators, but there were still many false positives or false negatives (i.e., situations in which behaviours were only annotated by one of the two annotators). The first author manually examined any other deviations in annotations and used this to redefine definitions or coding criteria (e.g., we changed the criteria for separating annotations of infant vocalisations and gestures, and more clearly defined the offset of an infant gesture which is important for starting the two-second response window, and better clarified some definitions). After improving the coding scheme, both annotators updated the pilot set accordingly. We achieved a satisfactory reliability score of κ = .75 (κ_{max} = .95; raw = .83), including unmatched annotations.

The last step involved an additional blind inter-annotator reliability check at the end. A random selection of seven videos was again double-coded by the first author. Overall, chance-corrected Cohen's kappa shows agreement of κ = .81 (κ_{max} = .94; raw = .87), including unmatched annotations, which is excellent. We also looked at agreement for each coding category. For infant behaviours, there was high agreement on the classification of infant vocalisations (κ = .70; κ_{max} = .85; raw = .97) and infant gestures (κ = .79; κ_{max} = .86; raw = .98). We found that the frequencies of the total number of observations per dyad were strongly correlated between the two annotators (r_s = .95, p < .01). This suggests that overall, identification and classification of infant vocalisations and gestures was strongly reliable. We also examined agreement on the categorisation of caregivers' verbal and nonverbal responses. For the binary variable indicating whether there was a response of any type, we find excellent agreement (κ

= .95, κ_{max} = .95; raw = .99). We also examined agreement on the different categories of caregivers' verbal responses (κ = .97; κ_{max} = .98; raw = .98), gestural responses (κ = .81; κ_{max} = 94; raw = .90), facial responses (κ = .74; κ_{max} = .87; raw = .97), and bodily responses (κ = .66; κ_{max} = .66; raw = .99) were all excellent. We find no more than three bodily responses in this set which is reflected in the lower maximum kappa. We reflect more on this in the discussion.

4.2.5. Statistical analyses

All analyses were carried out in *R* version 4.2.0 (R Core Team, 2022). First, we aimed to examine which infants' vocalisations, gestures, and bimodal behaviours and which caregivers' verbal, nonverbal, and multimodal response types occurred in this large, naturalistic dataset. In the first sections, we present descriptive statistics of all annotated behaviours to address this first aim. We calculated the total number of productions for each behaviour, the production range, and the proportion of participants who produced the behaviour at least once. Lastly, we examined which infant vocalisations and gestures and which caregiver verbal and nonverbal responses were often combined to form an infant bimodal behaviour or a caregiver multimodal response.

The second aim of this study was to examine whether different infant behaviours elicit different caregiver response rates and types. To examine this, we first fitted sets of logistic mixed-effects models using three binary outcomes indicating the presence or absence of a verbal, nonverbal, or multimodal response. All models were fitted with a random intercept for subjects using the *lme4* package version 1.1-31 (Bates et al., 2015) since we have multiple observations per dyad. In the first set of models, we used infant behaviour (vocalisation, gesture, or bimodal) as the predictor variable. Then, we also wanted to examine in more detail whether different gestures or different vocalisations also elicited different response rates. In the next set of models, we used infant gesture (index-finger pointing, whole-hand pointing, showing, giving, reaching, requesting, or other) or infant vocalisation (CV or non-CV) as the predictor variables. We used dummy coding with the category containing the largest number of

observations (infant behaviours: vocalisations; infant gestures: reaching; infant vocalisations: non-CV) as reference levels. Lastly, we present Chi-square test statistics to examine whether there was a relationship between infants' vocalisations and gestures and caregivers' verbal and gestural responses. We used the non-parametric Fisher's exact test to examine this for facial and bodily response types due to the low frequencies in these categories. The results indicate whether different infant behaviours elicited different response rates and types.

4.3. Results

4.3.1. Infant behaviours

In total, we annotated 2,036 infant behaviours of which 1,892 were infant vocalisations. All infants included in this study produced at least one non-CV vocalisation during the session. Of all vocalisations, only 55 were classified as CV, consisting of at least one syllable that did not only involve a glide or glottal. Approximately twenty percent of the infants in this study produced at least one CV vocalisation. Most of them produced only a few, although one infant produced thirteen instances during the session.

We annotated 207 infant gestures. Slightly more than half (53%) of the infants in our study produced at least one unimodal gesture. The most frequent infant gesture is reaching. We also annotated many instances of giving and index-finger pointing. Despite the young age, index-finger pointing was much more common than whole-hand pointing. The infants in this study did not spontaneously produce many showing or requesting gestures. The total frequencies of all infant behaviours are shown in Table 4.1.

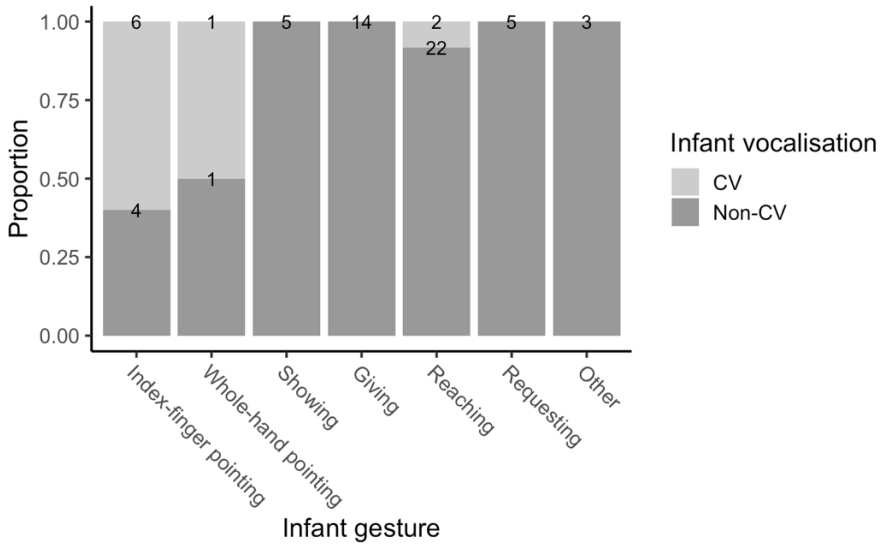
Table 4.1. Total frequencies of infant behaviours including production range and percentage of infants who produced the behaviour.

Infant behaviour	Frequency	Range	Percentage
Vocalisations			
CV vocalisation	55	0-13	19.7
Non-CV vocalisation	1837	1-54	100.0
Total	1892		100.0
Gestures			
Index-finger pointing	25	0-4	12.0
Whole-hand pointing	4	0-1	3.4
Showing	18	0-5	9.4
Giving	56	0-7	18.8
Reaching	78	0-6	30.8
Requesting	9	0-2	6.0
Other	17	0-7	4.3
Total	207		53.0

We also examined whether infants produced bimodal behaviours (i.e., vocal-gestural combinations). In total, only 63 infant behaviours were bimodal. At least one bimodal behaviour was produced by a quarter (25.6%) of the infants in this study. The gestures that were most often combined with a vocalisation were requesting (55.6% of instances were bimodal), whole-hand pointing (50%), and index-finger pointing (40%). Nine (14.3%) of the vocalisations in bimodal behaviours were CV vocalisations. In the full data set, only 2.8% of all vocalisations were classified as CV vocalisation. Therefore, infants tended to produce more CV vocalisations in bimodal behaviours compared to unimodal vocalisations. Figure 4.1 shows the proportions of each infant gesture combined with CV and non-CV vocalisations. This shows that

pointing gestures were relatively more often combined with CV vocalisations compared to other infant gestures.

Figure 4.1. Proportions of infants' bimodal gestures combined with CV and Non-CV vocalisations including raw counts.



4.3.2. Caregiver responses

In total, we annotated 2,036 infant behaviours of which 87.1% received a caregiver response of any type. A caregiver response could fall into multiple categories in the case of a multimodal response. Overall, caregivers produced more verbal than nonverbal responses. Table 4.2 shows verbal response frequencies, ranges, and percentages of caregivers who produced the response at least once. Most verbal responses are classified under contingent (e.g., talking about a toy that the infant is showing) or non-contingent (e.g., an affirmation) verbal responses, but we also annotated imitations of infant vocalisations and some onomatopoeias or sound effects (e.g., "broom broom").

Table 4.2. Total frequencies of caregivers' verbal responses including production range and percentage of caregivers who produced the response.

Caregiver response	Frequency	Range	Percentage
Contingent	670	0-27	93.2
Non-contingent	726	0-23	94.0
Infant imitation	127	0-11	47.0
Onomatopoeia	43	0-5	24.8
Total	1566		

We also annotated nonverbal responses shown in Table 4.3. We found many occurrences of manual gestures, such as object manipulations (e.g., riding the toy car around), and deictic gestures (e.g., showing or pointing). Deictic gestures were produced by approximately a third of the participants in this study. There were few representational gestures (e.g., demonstrating how the pop-up toy works without pressing the buttons) and only eleven gestures were classified as other, for example, waving or nodding. These categories were infrequent and therefore not included in further analyses. To a lesser extent, caregivers used their faces or bodies to respond to infants. Caregivers also frequently began to smile during or after the offset of the infant behaviour. Smiling as a response occurred in more than half of all caregivers in this study. There were not many occurrences of surprise, and we did not annotate any other facial expressions. The other non-facial and non-gestural bodily behaviours, such as a change of body orientation, did not occur frequently as a response to infants' vocalisations and gestures.

Table 4.3. Total frequencies of caregivers' nonverbal responses including production range and percentage of caregivers who produced the response.

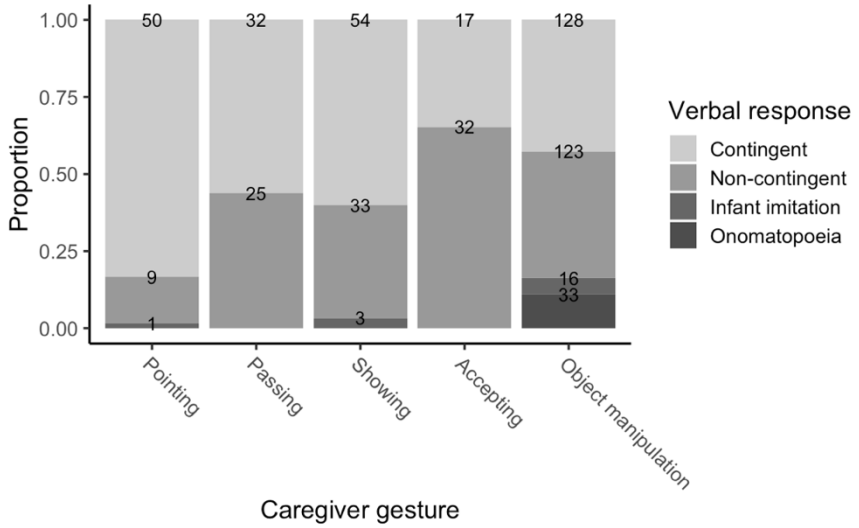
Caregiver response	Frequency	Range	Percentage
Gestural responses			
Pointing	61	0-5	29.1
Passing	74	0-6	36.8
Showing	101	0-22	36.8
Accepting	60	0-7	20.5
Representational	2	0-1	1.7
Object manipulation	388	0-13	83.8
Other	11	0-3	7.7
Total	697		
Facial responses			
Smile	168	0-10	65.0
Surprise	23	0-2	15.4
Total	191		
Bodily responses			
Leaning closer	26	0-3	17.9
Turning to infant	12	0-3	6.8
Turning to toy	6	0-2	4.3
Affective	36	0-3	20.5
Total	80		

Of all 1,774 caregiver responses, 39.7% were multimodal. This indicates that a verbal response was accompanied by a gestural, facial, and/or bodily response – at least partially overlapping in time. Most caregivers (94%) produced at least one multimodal response. There were only 209 unimodal nonverbal responses. Caregivers' gestural responses were most often multimodal out of all response categories (80.9%). Almost all pointing gestures produced by caregivers (98.4%) were multimodal. In contrast,

less than half (45%) of verbal responses were multimodal. Only onomatopoeias or sound effects show a high degree of multimodality (79.1%). Semantically contingent (i.e., follow-in) verbal responses are more often produced multimodally (50.6%) than non-contingent verbal responses (41.9%). Infant imitations were less often multimodal (21.3%). Lastly, other facial and bodily responses also occurred in high proportions of multimodal responses (58.6% and 78.8% respectively). Caregivers tend to combine nonverbal responses - mainly gestures and bodily behaviours - with verbal responses.

We further examined which caregivers' verbal and gestural responses often co-occurred. Figure 4.2 shows that, although frequently occurring in a multimodal response, onomatopoeias or sound effects are only combined with object manipulations. Caregivers' pointing gestures are more often combined with a semantically contingent response. Caregivers' showing and passing gestures are also more likely to be combined with a semantically contingent response. Caregivers rarely use infant vocal imitations in a multimodal response.

Figure 4.2. Proportions of caregivers' multimodal gestures combined with different verbal response types including raw counts.



4.3.3. Predicting response rates

We first compared response rates by calculating the proportions of responded-to behaviours for each infant vocalisation and gesture. Table 4.4 shows that infant gestures had higher overall response rates than infant vocalisations. All gestures examined in this study were highly frequently responded to.

Table 4.4. Raw counts and percentages of the total infant behaviours that elicited any type of response from caregivers.

Vocalisations	Count (% of total)	Gestures	Count (% of total)
CV	48 (87.3)	Index-finger pointing	24 (96.0)
Non-CV	1588 (86.4)	Whole-hand pointing	3 (75.0)
		Showing	17 (94.4)
		Giving	55 (98.2)
		Reaching	71 (91.0)
		Requesting	9 (100.0)
		Others	17 (100.0)

Next, we examined whether different categories of infant behaviours (vocalisations, gestures, and bimodal behaviours) elicited different proportions of verbal, nonverbal, and multimodal responses. Table 4.5 shows the percentages of each response category to each infant behaviour.

The results show that infant vocalisations received more caregiver verbal responses than nonverbal or multimodal responses. Infant gestures elicited many verbal and nonverbal responses. Lastly, infant bimodal behaviours elicited the highest percentages of verbal and multimodal responses.

Table 4.5. Raw counts and percentages of the total infant behaviour types that elicited any type of verbal, nonverbal, and multimodal response from caregivers.

	Caregiver response		
Infant behaviour	Verbal response	Nonverbal response	Multimodal response
Vocalisation	1394 (76.2)	771 (42.2)	586 (32.0)
Gesture	114 (79.2)	104 (72.2)	80 (56.6)
Bimodal	58 (92.1)	38 (60.3)	38 (60.3)

Three mixed-effects logistic regression models were fitted to test whether certain behaviours were statistically more likely to elicit 1) verbal responses, 2) nonverbal responses, and 3) multimodal responses. The first model indicates that infant bimodal behaviours are 6.42 times (95% CI [2.39, 17.23]) more likely to receive a verbal response than infant vocalisations. Infant vocalisations and infant gestures do not significantly differ in eliciting verbal response rates. The second model shows that infant bimodal behaviours are almost twice (95% CI [1.13, 3.43]) as likely to receive a nonverbal response, and infant gestures are almost 4 times (95% CI [2.63, 5.99]) more likely to receive a nonverbal response compared to infant vocalisations. The last model reveals that infant bimodal behaviours are 3.48 (95% CI [2.00, 6.06]) times more likely to elicit a multimodal response, and infant gestures are 2.96 (95% CI [2.02, 4.34]) times more likely to receive a multimodal response than infant vocalisations.

We also examined whether specific types of vocalisations or gestures are more likely to elicit specific caregiver response types. None of the models including infant vocalisation (CV or non-CV) as a predictor reached significance, indicating that the two types of vocalisations do not differ in eliciting verbal, nonverbal, or multimodal responses. The results indicate that the best predictor of a caregiver verbal response was infant index-finger pointing. Infant index-finger pointing was 14.82 times (95% CI [1.25, 175.08]) more likely to elicit a caregiver verbal response compared to infant

reaching. The other gestures did not significantly differ from infant reaching in eliciting verbal response rates. In contrast, infant giving gestures elicited higher nonverbal and multimodal response rates. Infant giving was 13.63 times (95% CI [3.10, 59.85]) more likely to receive a nonverbal response, and 5.11 times (95% CI [1.92, 13.65]) more likely to elicit a multimodal response compared to infant reaching. The other infant gestures did not significantly differ from reaching.

In sum, infant bimodal behaviours elicited more verbal and multimodal responses, while infant gestures elicited more nonverbal responses. Overall, infant vocalisations are less likely to elicit caregiver responses, but they received more verbal than nonverbal or multimodal responses. When examining the data in more detail, we found that verbal responses are more often elicited by infant index-finger pointing compared to other infant gestures. Lastly, we found that infant bimodal behaviours elicited more caregiver multimodal responses compared to unimodal vocalisations or gestures, although infant gestures elicited more multimodal responses than infant vocalisations.

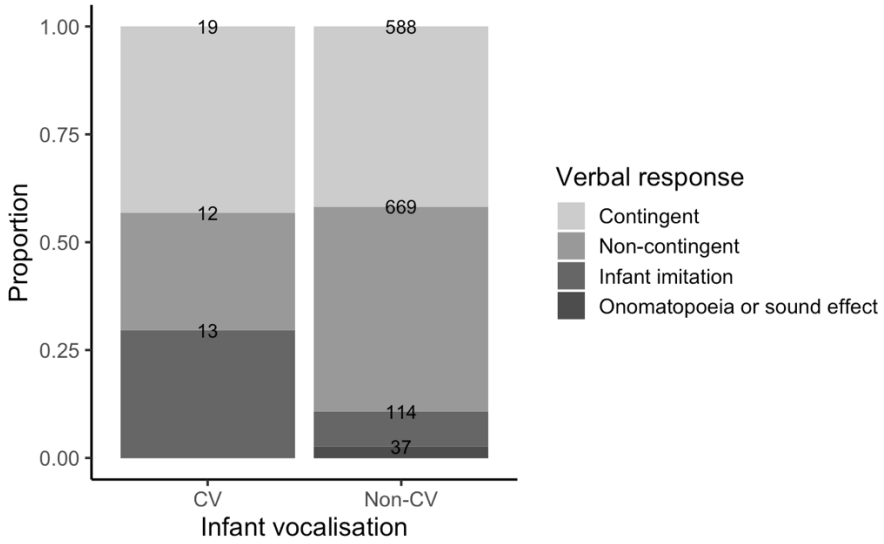
4.3.4. Predicting caregiver response types

We also aimed to examine whether different infant vocalisations and gestures elicited different caregiver verbal, gestural, facial, and bodily response types. In the case of a significant difference, we also present figures showing the proportions of each response category in response to different infant vocalisations and gestures to visually examine the differences.

Although we showed earlier that different infant vocalisations did not elicit different response rates, we do find that infant vocalisations elicited different verbal response types ($X^2 = 27.35$, $p < .001$). Figure 4.3 shows that caregivers verbally imitated CV vocalisations more often than non-CV vocalisations. Infant vocalisations did not elicit different gestural responses ($X^2 = 1.53$, $p = 0.82$), facial responses ($p = 0.06$), or bodily responses ($p = 1.0$). More data could be necessary for the latter two categories to detect

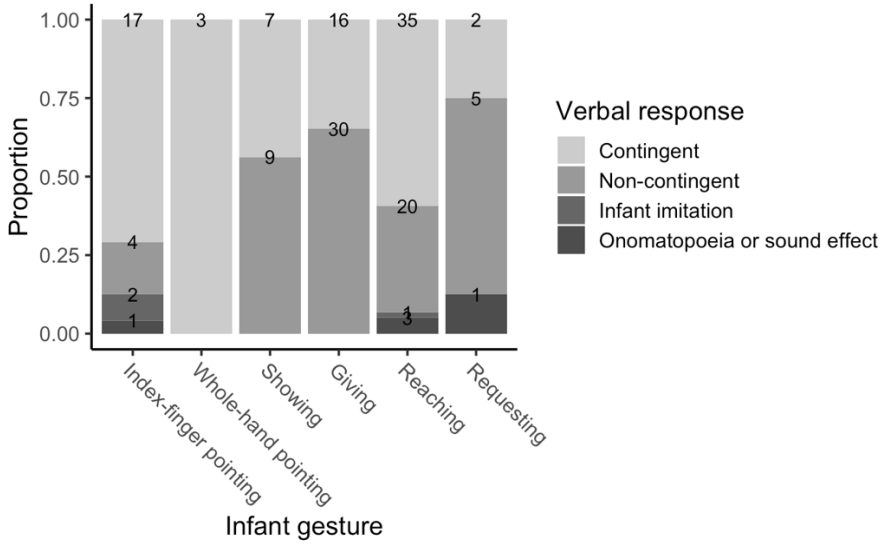
small effects, but these initial results suggest that vocalisations did not elicit different nonverbal responses.

Figure 4.3. Infants' vocalisations eliciting different proportions of verbal response types from caregivers including raw counts.



Next, we examined whether different infant gestures elicited different verbal and nonverbal response types. First, different infant gestures elicited different verbal responses ($X^2 = 35.20$, $p < .01$). Figure 4.4 shows that index-finger pointing, whole-hand pointing, and reaching elicited more contingent verbal responses (e.g., naming the object that the infant gesture was directed to), while all other gestures elicited more non-contingent verbal response (e.g., saying "oh nice" after the infant showed a toy).

Figure 4.4. Infants' gestures eliciting different proportions of verbal response types from caregivers including raw counts.



Infant gestures also elicited different gestural responses ($X^2 = 157.07, p < .001$). Figure 4.5 shows that infant giving was usually responded to by caregiver accepting. This is the most predictable caregiver gesture in response to any infant gesture. Infant requesting tended to elicit caregiver passing, and to a lesser extent caregiver pointing and caregiver showing. We see that infant pointing and requesting were the only types of gestures that elicited caregiver pointing in response. Infant showing elicited mostly caregiver accepting or object manipulations.

Lastly, we found that infant gestures elicited different facial expressions ($p < .01$) but not different bodily responses ($p = 0.31$). Again, more data may be needed in the latter category to detect any small effects. Nevertheless, the Fisher's exact test does indicate a difference in eliciting facial expressions. Figure 4.6 shows that infant showing and reaching only elicited smiles, while infant giving only elicited surprise. This is the first initial evidence that caregivers use different facial expressions to respond to different infant gestures, further suggesting that infants play an important role in the type of verbal and nonverbal language they receive.

Figure 4.5. Infants' gestures eliciting different proportions of gestural response types from caregivers including raw counts.

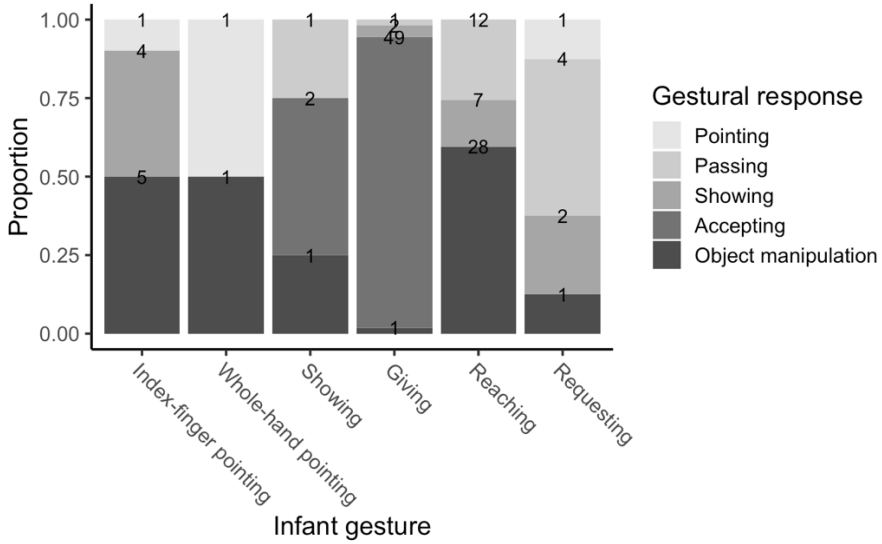
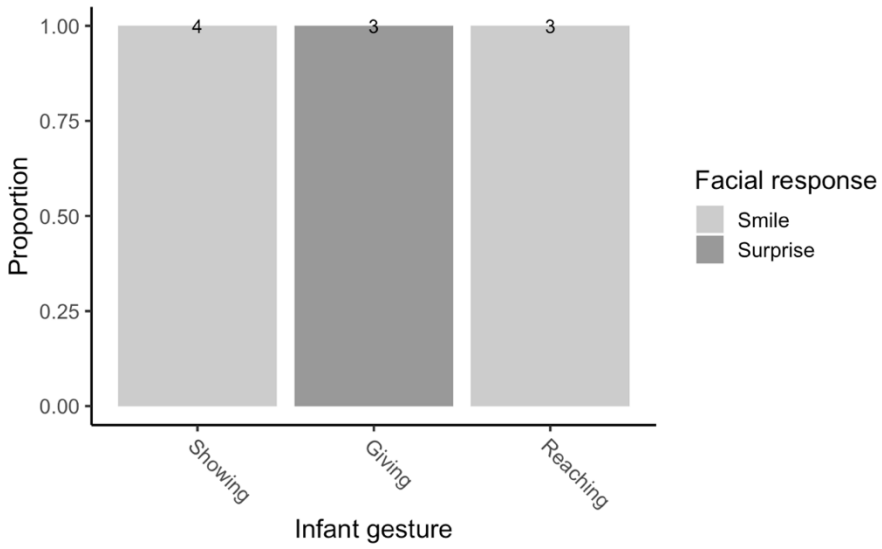


Figure 4.6. Infants' gestures eliciting different proportions of facial response types from caregivers including raw counts.



4.4. Discussion

In the present study, we developed a new coding scheme to annotate caregivers' verbal, nonverbal (i.e., gestural, facial, and bodily), and multimodal responses to infants' vocalisations and gestures. The first aim was to determine which infant vocalisations, gestures, and bimodal behaviours and which caregiver verbal, nonverbal, and multimodal responses occur in a large, naturalistic data set. The second aim was to assess whether different infant behaviours elicited specific caregiver response rates and types.

4.4.1. Variability in infant behaviours

Most infant behaviours were non-CV vocalisations, and they were produced at least once by all infants in our sample. This is to be expected, as this behaviour emerges early in development. We do find large individual variability across infants in the number of productions during the annotated six minutes. Many infants only produced up to ten non-CV vocalisations, while others produced over fifty. To a lesser extent, infants produced CV vocalisations. We only annotated 55 CV vocalisations in total, produced by only one-fifth of the infants in our sample. This is considerably less than the number of infants who produced CV vocalisations in the study by Donnellan et al. (2019), who found that 97% of 11-month-olds produced at least one CV vocalisation. The difference could be explained by the ages of the infants. The infants in their study were on average one month older, which is an important difference in the development of early vocalisations. Another explanation could be the duration of the interactions. In the study by Donnellan et al. (2019), caregiver-infant dyads were observed for 10 to 15 minutes which is much longer compared to the six minutes in the present study, giving them more time to produce less frequent behaviours.

The most frequent infant gestures in the current study were reaching, giving, and pointing. We found that reaching is the most frequent gesture overall, and it is produced by the largest group of infants in our study. Although less frequent than reaching, giving shows the largest individual variability across the infants in the current study. Although most infants did not produce these gestures even once, we

find some infants who produced them six or seven times in the annotated session. It may be surprising that we found more occurrences of index-finger pointing than showing, although the latter shows more individual variability. The present study was not designed to elicit pointing gestures. Yet, it was the third most frequent gesture among the infants in the current study. In the study by Donnellan et al. (2019), index-finger pointing was also produced by 21% of infants in their study, the largest group after giving and showing. Once infants have acquired index-finger pointing, they may rapidly start to produce many of them. Index-finger pointing has a clear deictic function which may be useful for prelinguistic infants to refine their early communicative bids.

We also examined which infant gestures were frequently combined with infant vocalisations, resulting in bimodal behaviours. Although the 10-month-olds in the current study did not produce many bimodal behaviours, the results show that infant requesting and pointing gestures have the largest proportions of bimodal productions. Approximately half of the total productions were combined with infant vocalisations. Pointing and requesting are among the later emerging infant gestures (Frank et al., 2021). The results of this study suggest that these later emerging gestures are more often combined with vocalisations. Infants only start producing bimodal behaviours later in development. Possibly, when infants are ready to produce pointing and requesting gestures, they are also ready to produce bimodal behaviours.

4.4.2. Caregivers' verbal, nonverbal, and multimodal responses

More than three-quarters of infant behaviours received a verbal response, making verbal responses the most common type of caregiver response. Caregivers tended to produce many contingent and non-contingent verbal responses. We also annotated some imitations of infant vocalisations, as well as onomatopoeias or sound effects. Regarding nonverbal responses, we annotated many gestural responses. We did not specifically elicit gestures during the caregiver-child interaction task, so these results suggest that caregivers tend to use many gestures naturally during early interactions with their infants. We annotated fewer facial and bodily responses. This could be due

to the criteria that we set for annotating a facial response: A facial expression was only annotated if the caregiver was not already showing the facial expression prior to the infant behaviour. We did find many smiles, compared to other facial responses, which is characteristic of infant-directed speech (Benders, 2013). We did not find many bodily responses which negatively affected the coding reliability of these responses. Nevertheless, there was high agreement between the two annotators on the few bodily responses that did occur. Therefore, it is likely that caregivers' vocalisations and gestures are more often used to respond to infants, while other non-gestural bodily behaviours naturally occur during interactions but may not be used specifically to respond to infants' vocalisations and gestures.

We also examined which verbal and nonverbal responses are often combined into multimodal responses. Of all verbal responses, approximately 40% were multimodal. Less than 10% of caregiver responses were nonverbal without any verbal component. This result agrees with the findings by Ger et al. (2018) who reported that approximately 40% of caregiver responses to infant index-finger pointing gestures were multimodal. We extended this finding to caregiver behaviours in response to all infant vocalisations and gestures. In a recent study, Chen et al. (2021) showed that caregivers often touched objects while naming them. Caregivers did this more often when the object was unfamiliar to the child, suggesting that they may do this to reduce referential ambiguity in learning contexts. Although we did not compare familiar to unfamiliar toys, our study extends this finding by showing that all caregivers' deictic gestures and object manipulations often co-occur with verbal responses. Caregivers' pointing gestures were most often multimodal. They were also more often combined with a semantically contingent verbal response compared to other gestures. This could be due to the deictic function of pointing. The contents of a semantically contingent response are often about one of the toys in the room. Caregivers frequently produced, for example, a labelling response while simultaneously pointing at the labelled object. In addition, caregivers' showing gestures were, although to a lesser extent, also more often combined with contingent verbal responses. A showing gesture also has a more deictic function than passing or accepting a toy. We also found that only object

manipulations are combined with onomatopoeias or sound effects. Caregivers tended to produce onomatopoeias or sound effects while playing with a toy, such as moving the toy car while saying "broom broom". This non-arbitrary connection could help the infant to connect the linguistic sound to the referent, aiding vocabulary development. Indeed, onomatopoeias are generally highly present in infant-directed speech (Motamedi et al., 2021). These visual cues, in the form of deictic gestures or object manipulations, are characteristic of infant-directed speech and may aid the infant's interpretation of unknown words (e.g., Baldwin et al., 1996; Gogate et al., 2000). Previous studies have predominantly examined caregivers' verbal responses, but the results of the current study suggest that it is important to also analyse gestural responses to gain a more complete picture of caregivers' responses.

4.4.3. Infant behaviours elicit specific responses

Our study found that different infant behaviours elicited different verbal, nonverbal, and multimodal responses. Different vocalisations by the infant elicited different types of caregiver verbal responses. Caregivers imitated infants' CV vocalisations more often than non-CV vocalisations, as previously found by Gros-Louis et al. (2006). Vocalisations did not elicit different nonverbal responses. Infant gestures, however, elicited different verbal, gestural, and facial responses. For example, infant index-finger pointing gestures elicited more semantically contingent verbal responses compared to the other infant gestures. Earlier research found that mothers provide more object labels to gestures compared to non-gestures (Olson & Masur, 2013) and that mothers are specifically more likely to provide object labels to infant index-finger pointing compared to infant reaching or infant vocalisations (Kishimoto et al., 2007; Wu & Gros-Louis, 2015). The results of the current study show that infant index-finger pointing gestures are often bimodal, and bimodal behaviours elicited the highest proportion of verbal responses. It could be possible that index-finger pointing has clear communicative intent, particularly when coordinated with a vocalisation, thereby eliciting more fine-tuned responses from caregivers. We also see many predictable patterns in the type of caregiver gestures that are elicited by infants. For example, infant giving gestures usually elicited caregiver accepting, while infant

showing, reaching, and requesting elicited higher rates of caregiver passing. Lastly, it appears that when caregivers accept toys from their infants, they tend to show a "mock surprised" facial expression. They did not show this expression in response to any other infant gesture. This suggests that even caregivers' facial expressions are dependent on infant behaviours. These results show the richness of multimodal communication between infants and their caregivers with highly predictable patterns within and across modalities.

Predictable infant behaviour and caregiver response patterns may highlight synchrony between infants and their caregivers. Synchrony can involve matching behaviours, such as smiling simultaneously. Previous studies showed that caregivers use more verbal than nonverbal responses to infant vocalisations, and infants' multimodal behaviours (looks combined with touch) elicited more caregivers' multimodal responses (looks combined with touch) (Gros-Louis et al., 2006; Yurkovic et al., 2021). This points to synchrony through matching modalities. That is, a co-occurrence of modality could suggest high synchrony between children and their caregivers (Leclère et al., 2014). Yet, these studies only examined infant vocalisations, looking behaviour, and touch. Our results add to these previous studies by showing that infant gestures were the best predictors of a nonverbal response, while infant bimodal behaviours elicited more multimodal responses compared to other behaviours. We did not find that infant vocalisations are the best predictors of a verbal response, because infant vocalisations elicited fewer caregiver responses in general. Nevertheless, of all caregiver responses that were elicited by infant vocalisations, the majority were indeed verbal. We thus extend previous findings regarding matched modalities to a wider set of behaviours, including infant vocalisations and gestures, as well as more caregiver verbal and nonverbal responses. It appears that caregivers tend to respond using the same modality as the infant behaviour.

The bidirectional approach taken in the current study poses new questions regarding the large variability in caregiver responsiveness. Previous studies have shown that caregivers individually differ in their verbal responsiveness which positively relates

to children's socio-cognitive outcomes, including language (e.g., McGillion et al., 2013; Olson & Masur, 2015; Wu & Gros-Louis, 2014). Caregivers differ in the number of verbal responses they tend to produce. For example, mothers are more likely to respond verbally than fathers (Flippin & Watson, 2011), and mothers of higher SES verbally respond more often than mothers of lower SES (e.g., Hart & Risley, 1995; McGillion et al., 2017; Vanormelingen & Gillis, 2016). Our results add to these previous studies by showing that responsiveness does not depend solely on caregiver characteristics. Infants influence the rates and types of responses they elicit from their caregivers by producing different types of vocalisations and gestures. More specifically, infants who produce many gestures and bimodal behaviours tend to elicit more caregiver responses than infants who predominantly produce vocalisations. This shifts the attention from individual behaviours to the dyad (Renzi et al., 2017; Chen et al., 2021). The question remains to what extent caregivers' responses in turn reinforce infants' behaviours. We should not study infant behaviours or caregiver responses in isolation to understand variability during early caregiver-child interactions, but rather we should examine the bidirectional effects that infants have on their caregivers and vice versa.

4.4.4. Limitations and future directions

The present study characterised infants' behaviours and caregivers' responses during free play. Caregivers likely use different types of cues in different learning environments. For example, some cues, such as body orientation, may become more important when objects are not in close proximity to the child and the caregiver. We may also expect infants to change their behaviours in different contexts. During book reading, infant index-finger pointing may be used more often than infant showing and passing, while the latter were most frequent during free play. Subsequently, caregivers' responsiveness will likely also be affected by the change in infant behaviours. Future studies can examine whether the predictable patterns between infant behaviours and caregiver responses remain stable, or whether we find differences in both infant behaviours, as well as caregiver response rates and types, across different learning environments.

The present study did not annotate nonverbal behaviours based on their contingency. In a recent study, Ger et al. (2018) have shown the important role played by contingent versus non-contingent responses that were measured both verbally and nonverbally. Future studies can examine the extent to which both verbal and nonverbal responses are semantically contingent on infants' behaviours. Previous studies have highlighted the important role played by intersensory redundancy in, for example, word learning (Gogate & Bahrick, 1998). Redundancy implies that there is some overlap in meaning between the verbal and nonverbal behaviour. Hence, in the case of redundancy, if the verbal response is contingent, the nonverbal behaviour should also be contingent. The question remains what proportion of multimodal responses typically contains redundant information.

We lastly want to acknowledge that the development of the coding scheme was highly informed by previous studies and empirical observations in the present study of western, educated, industrialised, rich, and democratic (WEIRD) caregivers. We recommend caution when using this coding scheme to annotate interactions of non-WEIRD dyads since the coding scheme reflects many cultural phenomena that may not be universal across cultures. By including onomatopoeias and nonverbal responses when examining caregivers' responsiveness, we already encompass more cultural diversity since caregivers from diverse cultures may produce different types of vocalisations, for example, Japanese mothers tend to produce far more onomatopoeias (Fernald & Morikawa, 1993), and Chinese mothers produce more pointing gestures compared to American mothers (So & Lim, 2012). However, it is always good to keep in mind that certain behaviours may occur in other cultures that are not included in the current coding scheme, but which do contribute importantly to caregiver responsiveness. It is important to validate a measuring instrument for specific populations.

4.5. Conclusions

This study provides an overview of caregivers' verbal, nonverbal, and multimodal responses to their 10-month-old infants' communicative behaviours in a large,

naturalistic data set. During free play, caregivers most often produced verbal responses, but 40% of those were multimodal. Caregivers often coordinated speech with manual and deictic gestures, and to a lesser extent with facial expressions and other bodily behaviours. Multimodal responses could be useful in learning contexts as they provide children with useful cues to disambiguate novel or unclear speech. We also examined whether different infant behaviours elicited different caregiver verbal, nonverbal, and multimodal responses. Infant bimodal (i.e., vocal-gestural combination) behaviours elicited high rates of verbal and multimodal responses, while unimodal gestures elicited high rates of nonverbal responses. We also found that different infant vocalisations elicited different verbal responses, while different infant gestures elicited different verbal, gestural, and facial responses. The results indicate that infants show large variability in the frequency and types of vocalisations and gestures they produce, which in turn affect when and how their caregivers respond. When examining caregiver-child interactions, analysing caregivers' verbal responses alone undermines the multimodal richness and bidirectionality of early communication.

Acknowledgements and data availability

We would like to thank Joyce van Zwet for her help with data annotation. We are also grateful to all families who participate in the YOUth study. YOUth is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). A complete listing of the study investigators and study management can be found at https://www.uu.nl/en/research/youth-cohort-study/about-us/who-isinvolved. YOUth investigators and management designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the YOUth study investigators or YOUth management. YOUth is a longitudinal study that aims to produce and safely store FAIR and high-quality data. The data can be accessed for both use and verification purposes upon request (see https://www.uu.nl/en/research/youth-cohort-study/data-access). All other

materials, detailed coding instructions, and *R* scripts are available online: https://osf.io/nvm54/.

References

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, *67*(6), 3135–3153. https://doi.org/10.2307/1131771

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 & Development*, *36*(4), 847–862. https://doi.org/10.1016/j.infbeh.2013.09.001

Bornstein, M. H., & Tamis-LeMonda, C. S. (1989). Maternal responsiveness and cognitive development in children. In *Maternal responsiveness: Characteristics and consequences* (pp. 49–61). Jossey-Bass.

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2019). Intention or attention before pointing: Do infants' early holdout gestures reflect evidence of a declarative motive? *Infancy*, *24*(2), 228–248. https://doi.org/10.1111/infa.12267

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/s030500090700829x

Chen, C., Houston, D. M., & Yu, C. (2021). Parent–child joint behaviors in novel object play create high-quality data for word learning. *Child Development*, *92*(5), 1889–1905. https://doi.org/10.1111/cdev.13620

Choi, B., Wei, R., & Rowe, M. L. (2021). Show, give, and point gestures across infancy differentially predict language development. *Developmental Psychology*, *57*(6), 851–862. https://doi.org/10.1037/dev0001195

Colonnesi, C., Stams, G. J. J. M., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366. https://doi.org/10.1016/j.dr.2010.10.001

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2019). Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants' vocalizations, gestures and word production. *Developmental Science*, *23*(1), 1–21. https://doi.org/10.1111/desc.12843

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. https://doi.org/10.2307/1166093

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur Communicative Development Inventories: User's guide and technical manual* (Second edition). Paul H. Brookes Publishing Co., Inc.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*(3), 637–656. https://doi.org/10.2307/1131208

Flippin, M., & Watson, L. R. (2011). Relationships between the responsiveness of fathers and mothers and the object play skills of children with autism spectrum disorders. *Journal of Early Intervention*, *33*(3), 220–234. https://doi.org/10.1177/1053815111427445

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank Project*. MIT Press. https://langcog.github.io/wordbank-book/

Ger, E., Altınok, N., Liszkowski, U., & Küntay, A. C. (2018). Development of infant pointing from 10 to 12 months: The role of relevant caregiver responsiveness. *Infancy*, *23*(5), 708–729. https://doi.org/10.1111/infa.12239

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*(4), 878–894. https://doi.org/10.1111/1467-8624.00197

Goldstein, M. H., Schwade, J. A., & Bornstein, M. H. (2009). The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child Development*, *80*(3), 636–644. https://doi.org/10.1111/j.1467-8624.2009.01287.x

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science*, *13*(1), 252–263. https://doi.org/10.1111/j.1467-7687.2009.00871.x

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*(6), 509–516. https://doi.org/10.1177/0165025406071914

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children* (pp. xxiii, 268). Paul H Brookes Publishing.

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

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. https://doi.org/10.1016/j.cogpsych.2010.08.002

Kishimoto, T., Shizawa, Y., Yasuda, J., Hinobayashi, T., & Minami, T. (2007). Do pointing gestures by infants provoke comments from adults? *Infant Behavior and Development*, *30*(4), 562–567. https://doi.org/10.1016/j.infbeh.2007.04.001

Kory Westlund, J. M., Dickens, L., Jeong, S., Harris, P. L., DeSteno, D., & Breazeal, C. L. (2017). Children use non-verbal cues to learn new words from robots as well as people. *International Journal of Child-Computer Interaction*, *13*, 1–9. https://doi.org/10.1016/j.ijcci.2017.04.001

Leclère, C., Viaux, S., Avril, M., Achard, C., Chetouani, M., Missonnier, S., & Cohen, D. (2014). Why synchrony matters during mother-child interactions: A systematic review. *PLOS ONE*, *9*(12), e113571. https://doi.org/10.1371/journal.pone.0113571

Liszkowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science*, *7*(3), 297–307. https://doi.org/10.1111/j.1467-7687.2004.00349.x

Masur, E. F. (1982). Mothers' responses to infants' object-related gestures: Influences on lexical development. *Journal of Child Language*, *9*(1), 23–30. https://doi.org/10.1017/S0305000900003585

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

McGillion, M. L., Pine, J. M., Herbert, J. S., & 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), 1122–1131. https://doi.org/10.1111/jcpp.12725

Motamedi, Y., Murgiano, M., Perniss, P., Wonnacott, E., Marshall, C., Goldin-Meadow, S., & Vigliocco, G. (2021). Linking language to sensory experience: Onomatopoeia in early language development. *Developmental Science*, *24*(3), e13066. https://doi.org/10.1111/desc.13066

Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating language in the real-world: The role of multimodal iconicity and indexicality. *Journal of Cognition*, *4*(1), 38. https://doi.org/10.5334/joc.113

Olson, J., & Masur, E. F. (2013). Mothers respond differently to infants' gestural versus nongestural communicative bids. *First Language*, *33*(4), 372–387. https://doi.org/10.1177/0142723713493346

Olson, J., & Masur, E. F. (2015). Mothers' labeling responses to infants' gestures predict vocabulary outcomes. *Journal of Child Language*, *42*(6), 1289–1311. https://doi.org/10.1017/S0305000914000828

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., van der Wal, J. J. M., Hulshoff Pol, H. E., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, *46*, 100868. https://doi.org/10.1016/j.dcn.2020.100868

Pearson, R. M., Heron, J., Melotti, R., Joinson, C., Stein, A., Ramchandani, P. G., & Evans, J. (2011). The association between observed non-verbal maternal responses at 12 months and later infant development at 18 months and IQ at 4 years: A longitudinal study. *Infant Behavior & Development*, *34*(4), 525–533. https://doi.org/10.1016/j.infbeh.2011.07.003

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Renzi, D. T., Romberg, A. R., Bolger, D. J., & Newman, R. S. (2017). Two minds are better than one: Cooperative communication as a new framework for understanding infant language learning. *Translational Issues in Psychological Science*, *3*, 19–33. https://doi.org/10.1037/tps0000088

Rowe, M. L., Wei, R., & Salo, V. C. (2022). Early gesture predicts later language development. In *Gesture in language: Development across the lifespan* (pp. 93–111). American Psychological Association. https://doi.org/10.1037/0000269-004

Ruddy, M. G., & Bornstein, M. H. (1982). Cognitive correlates of infant attention and maternal stimulation over the first year of life. *Child Development*, *53*(1), 183–188. https://doi.org/10.2307/1129651

Sloetjes, H., & Wittenburg, P. (2008, May). Annotation by Category: ELAN and ISO DCR. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. LREC 2008, Marrakech, Morocco.

So, W. C., & Lim, J. Y. (2012). "What is this?" Gesture as a potential cue to identify referents in discourse. *Applied Psycholinguistics*, *33*(2), 329–342. https://doi.org/10.1017/S0142716411000373

Tamis-LeMonda, C. S., & Bornstein, M. H. (2002). Maternal responsiveness and early language acquisition. In R. V. Kail & H. W. Reese (Eds.), *Advances in Child Development and Behavior* (Vol. 29, pp. 89–127). JAI.

Vanormelingen, L., & Gillis, S. (2016). The influence of socio-economic status on mothers' volubility and responsiveness in a monolingual Dutch-speaking sample. *First Language*, *36*(2), 140–156. https://doi.org/10.1177/0142723716639502

Verhagen, J., van den Berghe, R., Oudgenoeg-Paz, O., Küntay, A., & Leseman, P. (2019). Children's reliance on the non-verbal cues of a robot versus a human. *PLOS ONE*, *14*(12), e0217833. https://doi.org/10.1371/journal.pone.0217833

Vigliocco, G., Gu, Y., Donnellan, E., Grzyb, B., Brekelmans, G., Murgiano, M., Motamedi, Y., Brieke, R., & Perniss, P. (in preparation). *The Ecological Language (ECOLANG) corpus of multimodal dyadic communication*.

Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C. R., Milan Maillo, I., & Perniss, P. (2019). Onomatopoeias, gestures, actions and words in the input to children: How do caregivers use multimodal cues in their communication to children? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

von Eye, A., & von Eye, M. (2008). On the marginal dependency of Cohen's k. *European Psychologist*, *13*(4), 305–315. https://doi.org/10.1027/1016-9040.13.4.305

Wu, Z., & Gros-Louis, J. (2014). Infants' prelinguistic communicative acts and maternal responses: Relations to linguistic development. *First Language*, *34*(1), 72–90. https://doi.org/10.1177/0142723714521925

Wu, Z., & Gros-Louis, J. (2015). Caregivers provide more labeling responses to infants' pointing than to infants' object-directed vocalizations. *Journal of Child Language*, *42*(3), 538–561. https://doi.org/10.1017/S0305000914000221

Yurkovic, J. R., Kennedy, D. P., & Yu, C. (2021). Multimodal Behaviors from Children Elicit Parent Responses in Real-Time Social Interaction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/2qw0d216

Appendix A: Coding scheme

Infant vocalisations and gestures are annotated independently. Whenever the infant combines or overlaps a vocalisation and a gesture, annotate both behaviours and start measuring two seconds after the offset of the last behaviour for the caregiver's response. If the infant produces the same vocalisation or gesture again within these two seconds, we only annotate a caregiver response once, unless the caregiver gives two separate responses. If the infant produces a vocalisation and a gesture which overlap in time, we measure two seconds after the onset of the last behaviour (i.e., we annotate only one "no response" when the infant produces two pointing gestures within two seconds or when a vocalisation and gesture overlaps in time and their caregiver did not respond to either of them). This was done to allow caregivers enough time to respond to the behaviours, while some infants tend to repeat vocalisations continually.

Infant vocalisations. An infant vocalisation is any sound produced by the infant except vegetative sounds (e.g., hiccoughs or coughs) or distress sounds (e.g., crying or fussing). The two types of infant vocalisations are defined in Table 4.6.

Table 4.6. Coding scheme for infant vocalisations.

Infant vocalisation	**Definition**
Consonant-Vowel (CV)	At least one syllable contains a consonant-vowel sequence ("baba", "ma" etc.), excluding glides ("ja") and glottals ("ha") (e.g., Donnellan et al., 2019).
Non-CV	All other types of vocalisations except vegetative sounds (e.g., hiccoughs or coughs) or distress sounds (e.g., crying or fussing).

Infant gestures. We coded eight types of infant gestures, following previous studies (e.g., McGillion et al., 2013; Wu & Gros-Louis, 2014; Olson & Masur, 2015; Donnellan et al., 2019) and items defined in the CDIs. We initially distinguished between three types of pointing gestures: index-finger pointing, whole-hand pointing, and any other precursor of pointing. We did not annotate any occurrences of the latter category, so this was excluded from further analyses. The beginning of each gesture should be marked at the frame where the arm reached maximum extension, and the end should be marked at the frame where retraction of the arm begins. If the arm is extended within 200 milliseconds of the previous arm retraction, and the infant produces the same gesture again, this counts as a single occurrence (following Donnellan et al., 2019). All gestures are defined in Table 4.7.

Table 4.7. Coding scheme for infant gestures.

Infant gesture	**Definition**
Index-finger pointing	Infant extends their index-finger in the direction of the object or event while the other fingers are partially or entirely curled back while looking at an object or event. The arms must be extended, with empty hands, and the child should not lean forward or touch the object (following McGillion et al., 2017).
Whole-hand pointing	Infant extends a majority of fingers in the direction of the object or event (Donnellan et al., 2019).
Other pointing	Infant produces another precursor of pointing: e.g., fist extension or thump extension in the direction of the object or event.
Showing	Infant holds out an object with extended arm(s) towards the caregiver's face (adapted from Masur, 1982).
Giving	Infant holds out an object with either (or both) arms extended towards the caregiver's hands or in a way as to deliver the object to the caregiver.
Reaching	Infant extends either (or both) hands to get to an object out of reach. In this case, the infant may lean forward. Excluding movements that were the first phase of grasping an object already within reach (Masur, 1982). If the infant starts moving closer to the object to eventually grasp it, still annotate reaching before the infant started moving.
Requesting	Infant extends either (or both) hands to get an object out of reach without leaning towards it. The infant may open and close their hand.
Other	Infant produces any remaining conventional gestures, such as the infant raising their arms to initiate being picked up, waving, shrugging, nodding "yes" or "no", or blowing a kiss.

Coding caregiver responses

Responses should be temporally contingent and can occur during the infant behaviour or within two seconds after the offset of the infant behaviour (McGillion et al., 2013). A caregiver response occurs when the caregiver produces any type of verbal and/or non-verbal behaviour within this time frame. Each category is coded independently.

Table 4.8. Coding scheme for verbal responses.

Verbal response	**Definition**
Contingent	If its semantic content was related to the attentional state of the infant in the five seconds prior to the onset of the utterance (e.g., Donnellan et al., 2019). The utterance refers to an object that the child is holding, looking at, or has referenced by a gesture, or if the utterance is related to the activity in which the child is engaged (following e.g., McGillion et al., 2017).
Non-contingent	Other types of non-contingent responses, for example, affirmations ("good job!"), routines ("peek-a-boo"), directive acts ("Now get the doll", if the doll was not the infant's current focus of attention), and other questions not specifically about the current focus of attention ("What do you want next?").
Onomatopoeias or sound effects	Onomatopoeias ("knor knor") or other sound effects such as noises made with the mouth that represent a sound (e.g., snorting like a pig or making the noise of drinking something) (Vigliocco et al., in prep)
Infant imitation	Child vocal imitations.

Table 4.9. Coding scheme for gestural responses.

Gestural response	**Definition**
Pointing	Caregiver points towards an object or event.
Passing	Caregiver gives a toy to the infant.
Showing	Caregiver holds out an object with either (or both) hands to show the object to the infant without manipulating it.
Accepting	Caregiver accepts the toy that the infant is giving by grabbing it out of the infant's hand(s).
Representational	Caregiver shows the size, shape or how an object works without the object in hand (e.g., pretending to drink from a bottle or holding up hands to demonstrate the size) (Vigliocco et al., in prep).
Object manipulation	Caregiver depicts how to use an object or imitate how object moves or act (e.g., letting the baby doll drink from the bottle or showing how to use the pop-up toy) physically with the object (Vigliocco et al., in prep).
Other	Any remaining conventional gestures, such as picking the infant up, waving, shrugging, nodding "yes" or "no", or blowing a kiss.

Table 4.10. Coding scheme for facial responses.

Facial response	**Definition**
Smiling	Caregiver shows a happy expression, typically with the corners of the mouth turned up.
Surprise	Caregiver shows a surprised expression, typically with eyebrows raised and jaw dropped down.
Other	Caregiver shows any other facial expression different from neutral.

Table 4.11. Coding scheme for bodily responses.

Bodily response	**Definition**
Leaning closer	Caregiver comes closer in proximity to the infant.
Turning to infant	Caregiver turns their head or full body towards the infant when the caregiver was facing elsewhere.
Turning to toy	Caregiver turns their head or full body towards the toy when the caregiver was facing elsewhere.
Affective language	Caregiver shows positive affect towards the infant by cuddling, patting, kissing, or caressing the infant's cheek.
Other	Caregiver shows another clear non-gestural bodily reaction.

Chapter 5

The role of dyadic coordination of infants' behaviours and caregivers' verbal and multimodal responses in predicting vocabulary outcomes

Abstract

There is robust evidence that infants' gestures and vocalisations and caregivers' contingent responses predict later child vocabulary. Recent studies suggest that dyadic combinations of infants' behaviours and caregivers' responses are more robust predictors of children's vocabularies than these behaviours separately. This study aimed to compare the predictive value of 1) frequencies of infants' individual behaviours (vocalisations, points, and shows+gives) regardless of a caregiver response, 2) frequencies of infants' behaviours combined with caregivers' verbal responses, and 3) frequencies of infants' behaviours combined with caregivers' multimodal responses for children's vocabulary outcomes. We examined 114 caregiver-child dyads at 9–11 months and children's concurrent and longitudinal vocabulary outcomes at 2–4 years. We found that infants' points related to children's later receptive vocabularies, while infants' shows+gives gives (i.e., a combined category including shows and gives) related to children's later productive vocabularies – only when taking the instances that were combined with caregivers' multimodal responses into account. We also found that only dyadic shows+gives are related to infants' gesture repertoires. The results highlight the importance of examining dyadic combinations of infants' behaviours and caregivers' responses during interactions when examining relations to children's vocabulary development.

The full chapter is submitted to a journal:

van der Klis, A., Junge, C., Adriaans, F., & Kager, R. (submitted). The role of dyadic coordination of infants' behaviours and caregivers' verbal and multimodal responses in predicting vocabulary outcomes.

5.1. Introduction

Before the onset of their first words, infants start producing vocalisations and gestures to communicate. Gestures are predictors of children's language development (e.g., Brooks & Meltzoff, 2008; Rowe & Goldin-Meadow, 2009; Rowe et al., 2008). This is most often researched for deictic gestures. Deictic gestures include points (index-finger extensions), shows (holding out an object), and gives (passing on an object) which appear relatively early in children's development (e.g., Bates et al., 1979; Capone & McGregor, 2004). Children's points have been particularly well-studied in a broad age range of children and strongly predict their concurrent and longitudinal language outcomes (for meta-analyses, see Colonnesi et al., 2010; Kirk et al., 2022), although studies suggest that children's shows+gives (i.e., a combined category including both showing and giving gestures) may be precursors to points (Cameron-Faulkner et al., 2015; Choi et al., 2021) and are better predictors of children's later vocabulary skills than points when measured early from 10 to 12 months of age (Choi et al., 2021; Donnellan et al., 2019). The meanings of deictic gestures depend on the immediate context in which they are being used. Gestures and speech alike involve understanding the sign-referent relationship. In addition, experimental research shows that infants expect their caregivers to respond with gaze alternation and contingent comments (i.e., joint attention) to their shows+gives and points from at least 10 and 12 months of age respectively, which suggests that infants produce these gestures with the goal to share attention and interest with others (Boundy et al., 2019; Liszkowski et al., 2004). Therefore, infants' shows+gives and points could be related to their vocabulary outcomes through their shared reliance on symbolic meaning and/or through their ability to establish joint attention during interactions.

Language learning occurs in social contexts constructed by the infant and the caregiver (Renzi et al., 2017). Caregivers' immediate responses to infants' communicative behaviours are related to children's vocabulary development (e.g., Donnellan et al., 2019; McGillion et al., 2013; Olson & Masur, 2015; Wu & Gros-Louis, 2014). It has been hypothesised that infants immediately learn from the verbal

contents of the response if the response is semantically and temporally contingent on the infants' behaviour (see Tamis-LeMonda et al., 2014). Mothers verbally respond more often to infants' gestural than non-gestural communicative behaviours (Olson & Masur, 2013). More specifically, infants' points have been found to elicit more verbal labelling responses from adults (Kishimoto et al., 2007; Wu & Gros-Louis, 2015). For example, when the infant points at a doll, the caregiver may immediately respond with "That's a doll!". This could make it easier for the infant to match the phonological form "doll" onto the object. Olson and Masur (2015) demonstrated that caregivers' verbal labelling responses to infants' gestures completely mediated the relationship between infants' gestures and their vocabulary outcomes. Since infants' points have been found to elicit more labelling responses from caregivers, this could be one of the mechanisms through which infants' points are a robust predictor of their vocabulary outcomes. Studies do not typically find an effect of infants' vocalisations on children's vocabularies by themselves, while adults' semantically contingent responses directly elicited by infants' vocalisations – or dyadic combinations of infants' vocalisations and caregivers' contingent responses – can significantly predict children's vocabulary skills (Donnellan et al., 2019; Gros-Louis et al., 2014; Lopez et al., 2020; McGillion et al., 2013). Infants who gesture and vocalise more frequently could have better vocabulary outcomes because they create more word-learning opportunities for themselves by eliciting informative responses from caregivers. These studies highlight the importance of caregivers' responsiveness to infants' vocalisations and gestures in their ability to explain variation in children's vocabulary outcomes, but the influences of different types of caregivers' responses to different infants' behaviours have not yet been compared systematically. In this study, we will examine different combinations of dyadic behaviours (i.e., combined infant behaviours and caregiver responses) and examine their influences on children's concurrent and longitudinal vocabulary outcomes.

5.1.1. Multimodal language input

Previous studies have primarily focused on caregivers' verbal responses to infants' behaviours. However, there is ample evidence that caregivers' nonverbal cues can

contribute to word learning. Nonverbal cues in caregivers' responses towards infants, such as handing over a toy, pointing, or smiling, predict children's vocabulary outcomes (Pearson et al., 2011; Ruddy & Bornstein, 1982). For example, when the infant points at a rattle, the caregiver may pick it up and shake the rattle while saying "What a nice rattle!". This provides the infant with both a verbal and a visual cue as to what "rattle" refers to. Children can use gaze direction, body orientation, and index-finger pointing as cues to map words onto objects (e.g., Baldwin et al., 1996; Grassmann & Tomasello, 2010; Kory Westlund et al., 2017; Verhagen et al., 2019). Such nonverbal cues reduce any referential ambiguity in the language input. In addition, Ger et al. (2018) found that the proportion of caregivers' responses to infants' points that was multimodal (verbal + nonverbal) at 10 months positively predicted infants' points at 12 months, suggesting that caregivers' multimodal responses can reinforce infants' points. Similarly, Cameron-Faulkner et al. (2015) found a positive correlation between the frequency with which mothers acted on the target objects (e.g., by playing with it) after their infant produced shows+gives and their index-finger pointing frequency at 12 months. During such interactions, infants choose the object of interest, and their caregiver generally commented and/or acted upon it (e.g., by accepting and manipulating it), and then returned the object to the infant (Cameron-Faulkner et al., 2015). Lastly, a recent study suggests that caregivers' multimodal responses could increase the duration of children's looks at the toy (i.e., enhance attention) (Chen et al., 2021). These studies suggest that caregivers' multimodal responses could be useful in reinforcing infants' communicative behaviours during interactions, providing additional visual cues to reduce the referential ambiguity in the learning environment, and increasing infants' attention – subsequently facilitating children's vocabulary development.

Despite the robust evidence of the facilitative role of nonverbal behaviours, previous studies examining the link between caregivers' responses and children's vocabulary outcomes typically focused on caregivers' verbal responses (e.g., Donnellan et al., 2019; McGillion et al., 2013; Olson & Masur, 2015; Wu & Gros-Louis, 2014). Recently, Choi et al. (2021) annotated caregivers' verbal and nonverbal responses to

infants' shows+gives and points. They found that caregivers respond more often to infants' shows+gives than infants' points when the infants were 10 months of age, and only infants' show+gives at 10 months could predict children's vocabulary outcomes at 18 months. The authors hypothesised that shows+gives at 10 months were a better predictor of children's vocabularies because they elicited more responses from caregivers than points at this early age. They did not distinguish between caregivers' verbal and nonverbal responses; thus, it remains unclear whether they differentially relate to children's vocabulary outcomes. In Chapter 4, we showed that 9- to 11-month-old infants' gives elicited more multimodal (verbal + nonverbal) responses from caregivers, while infants' points elicited more verbal responses from caregivers. If infants' shows+gives are related to children's vocabulary outcomes due to the high rates of responses they elicit from caregivers at this early age, we would expect that caregivers' multimodal responses also contribute to the relationship between infants' shows+gives and children's vocabulary outcomes. To our knowledge, no prior studies have examined the effects of caregivers' multimodal responses on children's vocabulary outcomes.

5.1.2. Research aims

This study aimed to assess whether dyadic combinations of infants' vocalisations and gestures (shows+gives and points) and caregivers' verbal and multimodal responses during a free play session at 9–11 months of age improves the predictive value of these infants' behaviours for explaining variation in children's concurrent and later vocabulary at 2–4 years of age. To examine this, we compared the predictive value of three subsets of individual and dyadic behaviours: 1) frequencies of infants' individual behaviours (vocalisations, points, and shows+gives) regardless of a caregiver response, 2) frequencies of infants' behaviours combined with caregivers' verbal responses, and 3) frequencies of infants' behaviours combined with caregivers' multimodal responses. Based on the findings by Donnellan et al. (2019), we expected that infants' behaviours combined with caregivers' verbal responses are more robust predictors of children's vocabulary skills compared to all infants' behaviours regardless of a response. The influence of caregivers' multimodal responses on

children's vocabularies has not been studied directly. Based on previous studies, we expected that caregivers' multimodal responses can facilitate children's vocabulary development by reducing referential ambiguity in unclear or novel speech, by reinforcing infants' behaviours during interactions, and/or by increasing the infants' attention to toys (e.g., Baldwin et al., 1996; Cameron-Faulkner et al., 2015; Chen et al., 2019; Ger et al., 2018; Grassmann & Tomasello, 2010). Therefore, we hypothesised that infants' behaviours combined with caregivers' multimodal responses are better predictors of children's vocabulary outcomes compared to infants' individual behaviours. This would highlight the importance of examining dyadic behaviours during real-time interactions and adds to our understanding of the facilitative role of infants' vocalisations and gestures in their vocabulary development.

5.2. Methods

5.2.1. Participants

The data for this study are derived from YOUth, an ongoing longitudinal cohort study part of Utrecht University and University Medical Center Utrecht (Onland-Moret et al., 2020). YOUth has repeated measurements at regular intervals ("waves"). The sample for this study consisted of 114 infants (65 females) around 9–13 months of age (M = 10.7, SD = 0.9) during Wave 1 and their caregivers (90 mothers; 24 fathers). This is the same sample reported in Chapter 4, but we excluded three participants for the current study: one participant because the child has developmental language disorder, one participant because they were multilingual, and one participant because the child suffered from many ear infections during development. For the present study, we analysed these infants' concurrent and longitudinal vocabulary outcomes. When the children were followed up in Wave 2, they were around 2–4 years old (M = 2.7, SD = 0.5). There were approximately one to three years (M = 1.8, SD = 0.5) in between measurement waves, randomly varying per participant. Most caregivers in the sample completed a college or university degree (83.5%). The YOUth cohort study is carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki), and all caregivers have signed informed

consent. The study was approved by the medical ethical committee of the University Medical Center Utrecht (application number 14–7–221). Children received a picture book after participating in Wave 1 and a frog umbrella after participating in Wave 2.

5.2.2. Materials and procedure

Wave 1: Parent-child interaction

During the lab visit for Wave 1 when the infants were around 9–11 months of age, the infants and their caregivers were asked to sit on a blanket with a standard set of toys (a baby doll with a milk bottle, a green toy car, a Bumba pop-up toy, a sun-shaped rattle, a shape sorter, and a picture book) in a sparsely furnished room. Caregivers were instructed to play as if they were at home. Three Dome cameras that can be moved and zoomed in and out were placed around the blanket and one fixed camera filmed an overview of the scene. To capture sound, a fixed, standing (Sennheiser ME64/K6P condenser) microphone was positioned next to the blanket. After reading out the instructions, the research assistant would take place behind a screen, so they were completely out of view. The dyads completed five 3-minute sessions consecutively, resulting in a total of fifteen minutes. The tasks were completed in a fixed order: free play, playing with a shape sorter, reading a picture book, again free play, and cleaning up. For the present study, we analysed the two bouts of free play (six minutes in total) per dyad.

Wave 1: N_{YOUth}-CDI 1

After the lab visit, caregivers were instructed to fill out the N_{YOUth}-CDI 1, which is our adapted version of the N-CDI 1 and N-CDI-WG (Zink & Lejaegere, 2003). We replaced or removed 12 typical Flemish words with synonyms that are more common in Standard Dutch spoken in the Netherlands (e.g., we removed *mantel* from *jas(je) / mantel* (“coat”)). We also included the list containing 65 gestures and actions from the full-length N-CDI-WG (Zink & Lejaegere, 2002). This scale contains “early gestures” including the first communicative gestures (e.g., points, shows, and gives) and games and routines (e.g., playing peekaboo) and “late gestures” including actions with objects (e.g., eating with a spoon or fork) and pretending to be a caregiver (e.g.,

pretending to feed a doll). For infants in this age group, the gesture scale appears to be a relatively reliable measure of early vocabulary. Compared to the vocabulary scales for infants, the gesture scale does not suffer from floor effects, is less influenced by caregiver reporting biases, and is a better predictor of children's longitudinal vocabulary outcomes (see Chapter 3). For the 103 vocabulary items, caregivers were asked to check for each item whether their child *understands* or *speaks* the word — also when the child produces synonyms or pronunciation errors. The lists were fully digitised so caregivers could fill them out online. We scored the lists following the instructions of the manuals (Zink & Lejaegere, 2002, 2003). For the present study, we analysed infants' word comprehension (i.e., the number of items for which caregivers checked *understands* or *speaks*) and total gestures (i.e., the number of gestures for which caregivers checked *yes*, *sometimes*, and *often*). In total, 98 caregivers filled out the N$_{YOUth}$-CDI 1 during Wave 1.

Wave 2: N$_{YOUth}$-CDI 2

During Wave 2 when children were 2–4 years old, caregivers were asked to fill out a contraction of the short forms N-CDI 2A (16–30 months) and N-CDI 3 (30–37 months) (hereafter N$_{YOUth}$-CDI 2) (Zink & Lejaegere, 2003). The contraction resulted in a total number of 207 vocabulary items after removing the overlapping ones. The YOUth cohort study made the contraction because the second measurement wave covers a broad age range of children. Caregivers were asked to check the items that the child *speaks* — also in case the child produces synonyms or pronunciation errors. We replaced or removed 26 typical Flemish words with similar words that are more common in Standard Dutch spoken in the Netherlands (e.g., *bank* instead of *zetel/sofa* ("couch")). We analysed children's word production (i.e., the total number of items for which caregivers checked *speaks*). In total, 87 participants filled out the N$_{YOUth}$-CDI 2 during Wave 2.

Wave 2: PPVT-III-NL

During the lab visit for Wave 2, we administered the third version of the Dutch Peabody Picture Vocabulary Task (PPVT-III-NL), which is a lab-administered task

of receptive vocabulary (Schlichting, 2005). The task measures whether a person can match a spoken word to one of the four pictures (i.e., multiple choice). It is designed as a behavioural task in which the participant points to one of the images and the experimenter produces the target words and scores manually. For the YOUth cohort study, we developed a computerised version of the PPVT-III-NL. The experimenter runs a script on a computer with a touch screen where children are provided with recordings of the test items and four pictures on the screen. This controls for differences in speaker pronunciations and minimises the role of the experimenter. Children can use the touch screen to select one of the pictures after the target item has been presented. During the task, words become increasingly more complex. The PPVT-III-NL has a total of 204 items, divided into 17 sets of 12 items. The task terminates when the child makes nine or more errors in one set ("final set") (see Schlichting, 2005). The programme automatically subtracts the number of errors from the maximum score (which is the number of the final set * 12 items), resulting in the child's raw score. During the task, the child's caregiver was present in the back of the room out of the child's view. Caregivers were explicitly instructed not to help or communicate with the child. We excluded five participants from the total sample regarding analyses involving the PPVT-III-NL because the task was stopped prematurely.

5.2.3. Coding

All coding of infants' behaviours and caregivers' responses was done in ELAN version 6.0 (Sloetjes & Wittenburg, 2008) for the multimodal corpus reported in Chapter 4. A trained research assistant first annotated all infant vocalisations and gestures. Vocalisations were all sounds produced by the infants that were not vegetative or distress sounds. The gestures included in this study were (1) points (index-finger extensions), (2) shows (holding out an object with extended arm(s) directed at the caregiver's face), and (3) gives (holding out an object with extended arm(s) directed at the caregiver's hands or in a way as to deliver the object to the caregiver). Whole-hand points were excluded. We analysed these deictic gestures because previous studies have identified that these gestures elicit many responses

from caregivers and affect children's vocabulary outcomes (e.g., Choi et al., 2021; Donnellan et al., 2019; Wu & Gros-Louis, 2015). For the present study, we collapsed the showing and giving gestures into one category (shows+gives) given the heterogeneity in the definitions of these gestures (following Cameron-Faulkner et al., 2015). To assess inter-annotator reliabilities, we report chance-corrected modified Cohen's kappa (κ) using the built-in calculator in ELAN which is based on the EasyDIAg toolbox (Holle & Rein, 2015). A random selection of sixteen videos was double-coded by the first author. In this subset, chance-corrected kappa shows high agreement on infant vocalisations (κ = .92), infant points (κ = 1.0), and infant shows+gives (κ = .95).

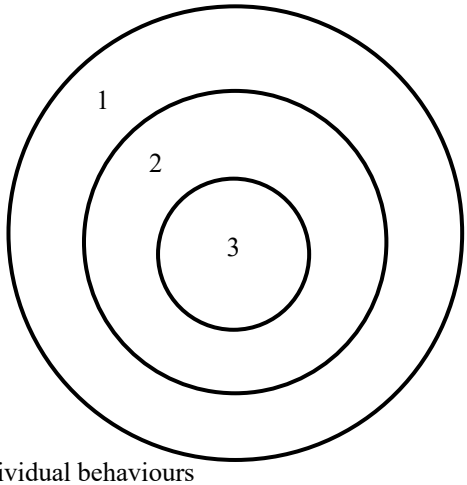
After the offset of the infant gesture or vocalisation, a period of two seconds was analysed for the caregiver response (following McGillion et al., 2013; Wu & Gros-Louis, 2014). The onset of the response had to occur during or within this two-second time frame. If not, the response was not considered temporally contingent and not included. In the case of a verbal response, we transcribed the utterance and annotated whether it was semantically contingent (i.e., a follow-in response related to the infant's focus of attention) or not. We assumed the object or activity was in the infant's focus of attention when the infant was vocalising while either holding the object or playing with the object, performing the activity, looking at the object, and/or gesturing towards the object. For the segmentation and classification of caregivers' semantically contingent verbal responses, we found high agreement (κ = .88) between the research assistant and the first author. The verbal response was coded multimodal when the verbal response was coordinated with at least one nonverbal cue, including gestural, facial, or other bodily responses. The nonverbal behaviour had to overlap the verbal response at least partially in time. We found high agreement on the classification of multimodal responses (κ = .84).

5.2.4. Analyses

All analyses were carried out in *R* version 4.2.0 (R Core Team, 2022). We contrasted frequencies of three subsets of infant and dyadic predictors on children's concurrent

and longitudinal vocabulary outcomes: 1) frequencies of infants' individual behaviours (vocalisations, points, and shows+gives) regardless of a caregiver response, 2) frequencies of infants' behaviours combined with caregivers' verbal responses, and 3) frequencies of infants' behaviours combined with caregivers' multimodal responses. Infant behaviours included vocalisations, points, and shows+gives. Verbal responses only included semantically and temporally contingent verbal responses (following Donnellan et al., 2019). The last subset included only the verbally contingent responses that were coordinated with a nonverbal cue (e.g., gestural, facial, or bodily behaviour) at least partially overlapping in time. For example, in subset 1, we analyse the frequency of infants' points. In subset 2, we only analyse the frequency of infants' points that elicited a verbal response from caregivers. In subset 3, we only analyse the frequency of infants' points that elicited a verbal response that is also combined with a nonverbal cue (i.e., multimodal response). Figure 5.1 depicts the three subsets of infant and dyadic behaviours. We examined which individual and/or dyadic behaviours predicted children's concurrent vocabulary comprehension and gesture repertoires (combined N-CDI 1 & N-CDI-WG) at Wave 1 and vocabulary comprehension (PPVT-III-NL) and vocabulary production (combined N-CDI 2A & N-CDI 3) at Wave 2. We fitted robust generalised linear models using the package *robustbase* version 0.95-0 (Maechler et al., 2022). We used raw scores for all vocabulary outcomes. We added children's ages in weeks and maternal education as a proxy for socio-economic status (SES) rated on a 9–point scale as control variables to the models. By adding children's ages during Wave 1 or Wave 2 as predictors to the models, all other predictors are independent of the effects of children's age. All continuous predictors were centred and scaled.

Figure 5.1. Three subsets of infant and dyadic behaviours used to predict children's vocabulary outcomes.



1 = infants' individual behaviours

2 = infants' behaviours combined with caregivers' verbal responses

3 = infants' behaviours combined with caregivers' multimodal responses

Analysing whether frequencies of caregivers' responses predicted children's vocabularies while controlling for frequencies of infants' behaviours – as would be a necessary step in a mediation analysis – was difficult due to the high multicollinearity across predictors. Frequencies of infant behaviours (e.g., frequencies of infants' points) and frequencies of caregiver responses to these infant behaviours (e.g., frequencies of caregivers' responses to infants' points) are by definition subsets of each other. Therefore, we compared the predictive value of individual behaviours versus dyadic behaviours in separate models instead. When we find that dyadic behaviours are stronger predictors of children's vocabularies compared to individual behaviours (by comparing the regression coefficients and statistical significance), this would suggest that dyadic behaviours are better predictors of children's vocabularies in this sample. Our approach thus uses the predictive value of infants' behaviours as a baseline against which to compare the predictive value of infants' behaviours combined with caregivers' verbal and multimodal responses to assess the relative contributions of dyadic versus individual behaviours.

5.3. Results

5.3.1. Descriptive statistics

First, we present the frequencies of infant behaviours during the six minutes of free play at Wave 1 from the multimodal corpus reported in Chapter 4. For the current study, we analysed the concurrent and longitudinal vocabulary outcomes of these infants during Wave 2. The descriptive statistics of all measures can be found in Table 5.1. Infants produced more vocalisations than gestures during the six minutes of free play. All infants in the sample produced at least one vocalisation, but they did not all produce points or shows+gives. On average, children produced fewer than one of these deictic gestures per session. However, some children in the sample spontaneously produced four pointing gestures, while some children produced up to eight shows+gives. This suggests there is individual variability across children in their productions of vocalisations and gestures.

Next, we examined caregivers' verbal and multimodal responses that were elicited by the infants' vocalisations and deictic gestures at Wave 1. The infants' behaviours were previously reported in the multimodal corpus in Chapter 4. The descriptive statistics of caregivers' responses to the infants' behaviours are presented in Table 5.2. For the current study, we assessed how many of the total infant behaviours elicited caregivers' semantically contingent verbal responses (i.e., verbal responses) (2) and caregivers' semantically contingent verbal responses that were coordinated with nonverbal behaviours (i.e., multimodal responses) (3). The remaining gestures elicited a different type of response from caregivers (i.e., a non-contingent verbal response or a gestural, facial, or bodily response that was not coordinated with a contingent verbal response), or no response at all.

Table 5.1. Descriptive statistics of the frequencies of infants' points, shows+gives, and vocalisations during the caregiver-child interaction task and raw scores of vocabulary outcomes.

	Wave	*M*	*SD*	Range
Infant behaviours				
Points	1	0.22	0.68	0 – 4
Shows+Gives	1	0.65	1.60	0 – 8
Vocalisations	1	16.32	11.09	1 – 54
Raw vocabulary scores				
NYOUth-CDI 1 Comprehension	1	40.59	23.10	0 – 99
NYOUth-CDI 1 Gestures	1	19.37	7.87	6 – 49
NYOUth-CDI 2 Production	2	152.46	42.19	23 – 207
PPVT-III-NL Comprehension	2	41.73	15.41	6 – 85

Table 5.2. Descriptive data of infant behaviours and the number of verbal, multimodal, or other/no responses they elicited from caregivers.

Infant behaviour	1. Total frequency	2. Verbal responses (%)	3. Multimodal responses (%)	4. Other or no responses (%)
Points	25	17 (68%)	7 (28%)	1 (4%)
Shows+Gives	74	23 (31%)	18 (24%)	33 (45%)
Vocalisations	1861	595 (32%)	291 (16%)	975 (52%)

More than two-thirds of the total infant points elicited semantically contingent verbal responses from caregivers. For example, caregivers named a toy after their infant had pointed at it. This proportion is considerably higher than the proportion elicited by

infant shows+gives or infant vocalisations. For the subset of caregivers' multimodal responses, the difference between infants' points and infants' shows+gives has become much smaller. Most caregiver responses elicited by infant shows+gives are multimodal (18/23, or 78%). For example, caregivers name a toy while accepting it from the infant who gives them the toy. This is not the case for the contingent responses elicited by infant points, where less than half were multimodal. Infant vocalisations elicited a relatively smaller proportion of multimodal responses from caregivers compared to infant gestures.

5.3.2. Predicting vocabulary outcomes with infant behaviours

First, we examined whether the frequencies of infants' behaviours (points, shows+gives, vocalisations) regardless of a caregiver response can predict children's concurrent and longitudinal vocabulary outcomes. The results are presented in Table 5.3.

First, the results show that the frequency of infant behaviours regardless of a caregiver response cannot predict infants' concurrent gestures or their later productive vocabularies measured using N_{YOUth}-CDIs with children's age and maternal education controlled. For infants, we find significant negative effects of maternal education (b = -8.57, SE = 2.85, $p < .01$) and the frequency of infants' points on the N_{YOUth}-CDI 1 comprehension (b = -3.43, SE = 1.11, $p < .01$). Lastly, infants' points are positively related to children's PPVT-III-NL receptive vocabulary outcomes years later (b = 2.48, SE = 0.64, $p < .001$).

Table 5.3. Robust regression coefficients of the models assessing influences of infant behaviours regardless of caregiver response on children's vocabulary outcomes.

	Wave 1		Wave 2	
	NYOUth-CDI 1 Comprehension	NYOUth-CDI 1 Gestures	NYOUth-CDI 2 Production	PPVT-III-NL
(Intercept)	39.35***	18.29***	155.53***	41.66***
Age in weeks	9.25***	4.09***	23.68***	10.61***
Maternal education	-8.57**	0.17	-2.53	2.04
Points	-3.43**	-0.21	-3.30	2.48***
Shows+Gives	-1.37	0.72	2.92	0.96
Vocalisations	0.69	0.26	-0.55	0.90
Observations	96	96	86	104
R^2	0.22	0.37	0.36	0.51
Adjusted R^2	0.18	0.34	0.32	0.49
Residual Std. Error	21.95 (df = 91)	5.45 (df = 91)	29.71 (df = 81)	10.08 (df = 99)

Note: * $p < .05$; ** $p < .01$; ***$p < .001$

5.3.3. Infant behaviours combined with caregivers' contingent verbal responses

In the next set of models, we only included the frequencies of infant behaviours that elicited a semantically contingent verbal response from caregivers. Apart from this, the reported models are identical to the previous models. The results are shown in Table 5.4.

First, we found that infant points are still negatively related to the NYOUth-CDI 1 comprehension ($b = -3.63$, $SE = 1.60$, $p = .03$) and positively related to children's later PPVT-III-NL receptive vocabularies ($b = 2.38$, $SE = 0.56$, $p < .001$) while controlling

for children's ages and maternal education. We also found that, when taking only infant behaviours that elicited semantically contingent verbal responses from caregivers into account, infants' shows+gives are positively related to their concurrent gesture repertoires ($b = 1.53$, $SE = 0.35$, $p < .001$). Infants' shows+gives regardless of a response do not predict infants' vocabulary skills as shown in Table 5.3, but higher frequencies of infants' shows+gives combined with caregivers' verbal responses have a positive effect on infants' concurrent gestures as shown in Table 5.4.

Table 5.4. Robust regression coefficients of the models assessing influences of infant behaviours that elicited verbal responses from caregivers on children's vocabulary outcomes.

	Wave 1		Wave 2	
	NYOUth-CDI 1 Comprehension	NYOUth-CDI 1 Gestures	NYOUth-CDI 2 Production	PPVT-III-NL
(Intercept)	39.78***	18.40***	155.56***	40.73***
Age in weeks	9.05***	4.13***	23.53***	10.96***
Maternal education	-7.82**	0.02	-2.10	1.87
Points	-3.63*	-0.44	-2.11	2.38***
Shows+Gives	1.01	1.53***	6.16	0.48
Vocalisations	2.31	0.48	3.27	0.73
Observations	96	96	86	104
R^2	0.21	0.43	0.36	0.52
Adjusted R^2	0.17	0.40	0.32	0.49
Residual Std. Error	21.75 (df = 91)	5.22 (df = 91)	30.86 (df = 81)	9.89 (df = 99)

Note: * $p < .05$; ** $p < .01$; ***$p < .001$

5.3.4. Infant behaviours combined with caregivers' multimodal responses

In the last set of models, we only included the subset of infants' behaviours which elicited caregivers' contingent verbal responses coordinated with at least one nonverbal cue (i.e., a multimodal response). The results are reported in Table 5.5.

Table 5.5. Robust regression coefficients of the models assessing influences of infant behaviours that elicited multimodal responses from caregivers on children's vocabulary outcomes.

	Wave 1		Wave 2	
	N_{YOUth}-CDI 1 Comprehension	N_{YOUth}-CDI 1 Gestures	N_{YOUth}-CDI 2 Production	PPVT-III-NL
(Intercept)	39.65***	18.41***	154.82***	40.80***
Age in weeks	8.65***	4.17***	23.15***	11.03***
Maternal education	-7.04**	0.09	-0.01	1.44
Points	-0.76	-0.31	-8.25	1.40
Shows+Gives	1.33	1.20**	5.98*	0.61
Vocalisations	1.94	0.66	4.43	0.33
Observations	96	96	86	104
R^2	0.19	0.41	0.40	0.49
Adjusted R^2	0.15	0.38	0.36	0.46
Residual Std. Error	22.25 (df = 91)	5.41 (df = 91)	29.98 (df = 81)	10.77 (df = 99)

Note: * $p < .05$; ** $p < .01$; ***$p < .001$

First, we found that infants' shows+gives that elicited contingent multimodal responses from caregivers still positively related to children's concurrent gesture repertoires ($b = 1.20$, $SE = 0.38$, $p < .01$). We also found that, when only taking infant

behaviours that elicited a contingent multimodal response from caregivers into account, infant shows+gives are positively related to children's later productive vocabularies measured by the NYOUth-CDI 2 (b = 5.98, SE = 2.33, p = .01). Lastly, we found that when including only the number of infant points that elicited a contingent multimodal response from caregivers, this behaviour does not significantly relate to children's NYOUth-CDI 1 comprehension or later PPVT-III-NL outcomes anymore. The results are discussed below.

5.4. Discussion

The goal of this study was to assess whether infants' vocalisations and gestures combined with caregivers' verbal and multimodal responses are better predictors of children's vocabulary outcomes than infants' individual behaviours separately. To examine this, we contrasted models with three different types of predictors on children's concurrent and longitudinal vocabulary outcomes: 1) frequencies of infants' individual behaviours (vocalisations, points, and shows+gives) regardless of a caregiver response, 2) frequencies of infants' behaviours combined with caregivers' verbal responses, and 3) frequencies of infants' behaviours combined with caregivers' multimodal responses. The results of this study improve our understanding of the facilitative role of caregivers' contingent and multimodal responses in children's vocabulary development. We first discuss the effects of infant and dyadic points, then shows+gives, and lastly vocalisations on children's concurrent and longitudinal vocabulary outcomes and concurrent gesture repertoires.

5.4.1. Infant points predict long-term vocabulary outcomes

First, when examining all infants' behaviours regardless of caregivers' responses, we found that only infants' points are related to their receptive vocabulary skills measured several years later. The finding that infants' points are a robust predictor of their longitudinal vocabulary outcomes agrees with previous studies, as infants' points have often been found to predict children's concurrent and longitudinal vocabulary outcomes (for meta-analyses, see Colonnesi et al., 2010; Kirk et al., 2022). In the current study, we did not find a concurrent relation between infant points and

children's gesture repertories. The cross-linguistic mean age of acquisition for the pointing gesture is 10.4 months of age (Frank et al., 2021). Therefore, it is likely that the infants in our study who pointed have only started doing so recently. When infants' points are facilitative because they tend to elicit semantically contingent responses from caregivers, we expect that it would take some time before the facilitative effect on children's vocabulary size can manifest. The two meta-analyses revealed that the relationship between infants' points and their concurrent and longitudinal vocabulary skills becomes stronger when pointing is measured later in children's development (Colonnesi et al., 2010; Kirk et al., 2022). The results of our study suggest that, although infants' points measured around 9–11 months of age do not predict their concurrent gesture repertoires, they are predictive of children's long-term receptive vocabulary outcomes.

We found that maternal education and the frequency of infants' points are negatively related to infants' concurrent word comprehension skills. Each of these findings seems contradictory at first because 1) maternal education has often been positively associated with children's vocabularies (e.g., Feldman et al., 2000; Fenson et al., 2007; Hoff, 2003) and 2) infants' points have been found to positively relate to children's concurrent and longitudinal vocabularies in many studies, particularly word comprehension (Colonnesi et al., 2010; Kirk et al., 2022). However, previous studies examining the effects of maternal education often report a negative effect of maternal education (or SES) on infants' word comprehension skills reported by their caregivers (e.g., Feldman et al., 2000; Reese & Read, 2000). This effect is likely caused by a caregiver reporting bias, where lower SES caregivers may overreport their infants' vocabularies because they think larger vocabularies are more desirable or higher SES caregivers may underreport their children's vocabularies because they underestimate their infants' skills, resulting in a negative SES effect. In contrast, previous studies reported positive links between infants' gestures and maternal education (Rowe & Goldin-Meadow, 2009). If children from higher SES families tend to produce more pointing gestures, this could result in a negative effect of infants' points on their concurrent word comprehension outcomes if the results are negatively influenced by

a reporting bias driven by caregivers' SES. The finding asks for further research on the effects of infants' points on their receptive vocabularies in a more diverse sample of children.

When taking only infants' points combined with caregivers' verbal responses into account, the variable remains a significant predictor of children's later receptive vocabularies. More than half of the infants' points elicited contingent responses from caregivers, so these predictors (infant points regardless of a caregiver response and infant points that elicited verbal responses from caregivers) are rather similar. Yet, the fact that losing over thirty per cent of data points does not impact the predictive value of points suggests that either 1) infants' points are a very robust predictor of children's long-term receptive vocabularies and/or 2) the relationship between infants' points and their vocabulary outcomes is driven by the contingent verbal responses that infants' points tend to elicit from caregivers. In support of the latter hypothesis, previous studies found that infants' points specifically tend to elicit more semantically contingent or labelling responses from caregivers (Kishimoto et al., 2007; Wu & Gros-Louis, 2015, also see Chapter 4). There is a frequent pattern of an infant pointing at a toy and the caregiver immediately naming that toy after the infant expressed interest in it. It could be possible that hearing frequent object labelling, particularly for objects that infants are interested in, improves children's word comprehension skills. When infants produce many pointing gestures, they create many word-learning opportunities for themselves. It is difficult to tease apart the effects of points from the effects of infants' points combined with caregivers' contingent responses in naturalistic caregiver-child interaction data, because infants' points tend to elicit high rates of verbal responses, making the two predictors rather similar.

5.4.2. Dyadic shows+gives predict gestures and vocabulary size

Recently, Choi et al. (2021) found that for 10-month-olds, shows+gives is a better predictor of children's later vocabulary skills than points, but shows+gives was not a significant predictor anymore by 12 months. Only from 14 months onwards, infants' points became a significant predictor. In contrast to Choi et al. (2021), we did not find

significant effects of infants' shows+gives on children's concurrent or longitudinal vocabulary skills when including all infant behaviours regardless of caregivers' responses. The children included in the present study are of a broader age range (9–11 months), and gestures develop rapidly during this period (e.g., Frank et al., 2021). In addition, the children in our study are from predominantly high SES backgrounds while the children in Choi et al. (2021) were from diverse backgrounds. Infants from higher SES families generally start producing points earlier (Rowe & Goldin-Meadow, 2009) which could speed up their vocabulary development. The high SES infants in our study could show faster progression in their gesture development, and therefore points could have already gained more predictive value than shows+gives. Nevertheless, the results of our study do not show that 10-month-old infants' shows+gives significantly correlate with any of the concurrent or longitudinal vocabulary measures.

We also examined whether infants' shows+gives that elicited contingent verbal responses from caregivers are better predictors of children's vocabulary outcomes compared to infants' shows+gives separately. We found that infants' shows+gives combined with caregivers' verbal and multimodal responses are significant predictors of children's concurrent gesture repertoires. Infants' gesture repertoires positively influence children's longitudinal vocabulary skills (e.g., Fenson et al., 2007), but the influence of infants' shows+gives combined with caregivers' verbal responses may not have been large enough to show a facilitative effect on children's longitudinal vocabulary outcomes. When only taking infants' shows+gives combined with caregivers' multimodal responses into account, we found that shows+gives do significantly relate to children's later productive vocabularies. Infants' gives tend to elicit higher rates of multimodal responses from caregivers compared to other gestures (see Chapter 4). For example, the caregiver accepts the object while talking about it. The combination of talking about an object while touching the object could boost children's word-learning abilities. Recently, Chen et al. (2021) showed that both naming and touching by the caregiver increase the duration of children's looks at the toy. By talking about the object and interacting with it physically, this type of response

is particularly likely to establish joint attention. Such responses also satisfy the infant's desire to share attention and interest with others (Boundy et al., 2019). In addition, naming combined with a nonverbal cue towards or with the object could reduce the referential ambiguity in the learning context (Baldwin et al., 1996; Grassmann & Tomasello, 2010; Kory Westlund et al., 2017; Verhagen et al., 2019). Since infants' show+gives tend to elicit high rates of multimodal responses from caregivers, they could facilitate children's vocabulary outcomes by reducing referential ambiguity in unclear or novel speech, by clearly establishing joint attention between the infant and the caregiver, and/or by enhancing the infants' attention to the toy.

5.4.3. Infant vocalisations do not predict vocabulary

We did not find any effects of infants' prelinguistic vocalisations on their vocabulary outcomes. In a previous study, 11-month-old infants' gaze-coordinated and responded-to vocalisations were the best predictors of children's vocabularies (Donnellan et al., 2019). Less than a quarter of all infant vocalisations were gaze-coordinated in the study by Donnellan et al. (2019). In our study, individual infants' vocalisations regardless of a caregiver response versus infants' vocalisations combined with a caregiver response did not influence children's vocabulary outcomes either way. It could be possible that we did not find an effect because we did not measure infants' gaze direction during vocalising. While gaze-coordination results in higher response rates from caregivers (Donnellan et al., 2019), measuring infants' vocalisations combined with caregivers' responses regardless of gaze did not have any predictive value for children's vocabulary skills in our study. It could be possible that measuring gaze to determine children's communicative intent is less important for infant gestures than infant vocalisations. Infants produce far more vocalisations than gestures, and it may be more difficult for caregivers to determine whether children produce vocalisations with the intention to communicate or to determine their communicative goal. Another explanation is that there are fast developmental changes across children within this age group. The children in the study by Donnellan et al. (2019) were slightly older and produced more advanced types of Consonant-Vowel

(CV) vocalisations (i.e., canonical babbles) compared to non-CV vocalisations. Although Donnellan et al. (2019) did not find differences between CV and non-CV vocalisations, the grouped variable including all types of vocalisations could gain more predictive value when this variable includes more canonical babbles. The predictive value of infants' vocalisations and gestures could therefore change continuously across children's development.

5.4.4. Limitations and future directions

The current study has several limitations that should be addressed in future studies. Even though the sample in our study is large, it overrepresents highly educated, white, western caregivers. Caregivers' education is known to affect infants' gestures, caregivers' contingent speech, and children's vocabulary outcomes (e.g., Hoff, 2003; McGillion et al., 2017; Rowe & Goldin-Meadow, 2009). Therefore, it is important to verify our findings in a more socio-economically diverse sample, especially concerning the finding that maternal education and the frequency of infants' points negatively affect infants' vocabulary comprehension. In addition, caregivers from different cultures may not all be equally talkative to their infants (see Cristia et al., 2017) which could also influence their response rates to infants' prelinguistic behaviours. Future studies should determine whether infants' points and infants' shows+gives still relate to children's vocabulary outcomes in diverse cultures where adults are less responsive to infants. If infants' gestures still relate to their vocabulary outcomes in these cultures, this suggests that mechanisms other than learning from caregivers' responses are also at play.

In addition, we only examined one setting (i.e., free play) at one point in time (i.e., at 9–11 months of age) in the current study. Infants and caregivers are both likely to change their behaviours depending in the environment. For example, we would expect infants to use more referential gestures when objects are further out of reach. During book reading, infants could be more likely to use points to refer to different pictures on the pages, while infants' shows+gives were most common during free play with a set of toys. Subsequently, caregivers' responsiveness will be influenced by the

changes in infants' behaviours (see Chapter 4). In addition, we may expect the influence of different dyadic combinations of behaviours on children's vocabulary outcomes to change across children's development. As shown by Choi et al. (2021), the predictive value of infants' points and infants' shows+gives for their vocabulary outcomes changes from 10 to 14 months. Caregivers' responsiveness also changes over time. For example, Choi et al. (2021) showed that caregivers respond more often to 10-month-old shows+gives than points. However, by 14 months, caregivers respond to both deictic gestures equally often. In sum, dyadic behaviours are expected to change across different settings and children's developmental stages – differentially predicting children's vocabulary outcomes.

5.5. Conclusions

A significant contribution of our study is that we have provided evidence for the predictive value of infants' deictic gestures, including points and show+gives, measured at a very young age (9–11 months of age) on children's long-term vocabulary outcomes (2–4 years of age). Although measured early, the facilitating effects of infants' gestures on their vocabulary development remain present all through the crucial years of rapid vocabulary development. Then, we showed that the predictive value of gestures can differ depending on the types of caregivers' responses they tend to elicit. The results suggest that infants' points tend to elicit verbal responses from caregivers which facilitate children's word comprehension skills, while infants' shows+gives tend to elicit multimodal responses from caregivers which facilitate children's word production skills. Importantly, only infants' shows+gives that were combined with caregivers' multimodal responses significantly predicted later child vocabulary. The results suggest that specific dyadic combinations of infants' gestures and caregivers' verbal and multimodal responses are more robust predictors of children's long-term vocabulary outcomes than infants' gestures separately.

Acknowledgements and data availability

YOUth is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). A complete listing of the study investigators and study management can be found at https://www.uu.nl/en/research/youth-cohort-study/about-us/who-isinvolved. YOUth investigators and management designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the YOUth study investigators or YOUth management. YOUth is a longitudinal study that aims to produce and safely store FAIR and high-quality data. The data can be accessed for both use and verification purposes upon request (see https://www.uu.nl/en/research/youth-cohort-study/data-access). The preregistration, *R* script, and other materials can be found online: https://osf.io/zxnqd/.

References

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, *67*(6), 3135–3153. https://doi.org/10.2307/1131771

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy*. Academic Press. https://doi.org/10.1016/C2013-0-10341-8

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2019). Intention or attention before pointing: Do infants' early holdout gestures reflect evidence of a declarative motive? *Infancy*, *24*(2), 228–248. https://doi.org/10.1111/infa.12267

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/s030500090700829x

Cameron-Faulkner, T., Theakston, A., Lieven, E., & Tomasello, M. (2015). The relationship between infant holdout and gives, and pointing. *Infancy*, *20*, 576–586. https://doi.org/10.1111/infa.12085

Capone, N. C., & McGregor, K. K. (2004). Gesture development. *Journal of Speech, Language, and Hearing Research*, *47*(1), 173–186. https://doi.org/10.1044/1092-4388(2004/015)

Chen, C., Houston, D. M., & Yu, C. (2021). Parent–child joint behaviors in novel object play create high-quality data for word learning. *Child Development*, *92*(5), 1889–1905. https://doi.org/10.1111/cdev.13620

Choi, B., Wei, R., & Rowe, M. L. (2021). Show, give, and point gestures across infancy differentially predict language development. *Developmental Psychology*, *57*(6), 851–862. https://doi.org/10.1037/dev0001195

Colonnesi, C., Stams, G. J. J. M., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366. https://doi.org/10.1016/j.dr.2010.10.001

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, *90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2019). Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants' vocalizations, gestures and word production. *Developmental Science*, *23*(1), 1–21. https://doi.org/10.1111/desc.12843

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*(2), 310–322. https://doi.org/10.1111/1467-8624.00146

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur Communicative Development Inventories: User's guide and technical manual* (Second edition). Paul H. Brookes Publishing Co., Inc.

Ger, E., Altınok, N., Liszkowski, U., & Küntay, A. C. (2018). Development of infant pointing from 10 to 12 months: The role of relevant caregiver responsiveness. *Infancy*, *23*(5), 708–729. https://doi.org/10.1111/infa.12239

Goldstein, M. H., Schwade, J. A., & Bornstein, M. H. (2009). The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child Development*, *80*(3), 636–644. https://doi.org/10.1111/j.1467-8624.2009.01287.x

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science*, *13*(1), 252–263. https://doi.org/10.1111/j.1467-7687.2009.00871.x

Gros-Louis, J., West, M. J., & King, A. P. (2014). Maternal responsiveness and the development of directed vocalizing in social interactions. *Infancy*, *19*(4), 385–408. https://doi.org/10.1111/infa.12054

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

Kirk, E., Donnelly, S., Furman, R., Warmington, M., Glanville, J., & Eggleston, A. (2022). The relationship between infant pointing and language development: A meta-analytic review. *Developmental Review*, *64*, 101023. https://doi.org/10.1016/j.dr.2022.101023

Kishimoto, T., Shizawa, Y., Yasuda, J., Hinobayashi, T., & Minami, T. (2007). Do pointing gestures by infants provoke comments from adults? *Infant Behavior & Development*, *30*(4), 562–567. https://doi.org/10.1016/j.infbeh.2007.04.001

Kory Westlund, J. M., Dickens, L., Jeong, S., Harris, P. L., DeSteno, D., & Breazeal, C. L. (2017). Children use non-verbal cues to learn new words from robots as well as people. *International Journal of Child-Computer Interaction*, *13*, 1–9. https://doi.org/10.1016/j.ijcci.2017.04.001

Liszkowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science*, *7*(3), 297–307. https://doi.org/10.1111/j.1467-7687.2004.00349.x

Lopez, L. D., Walle, E. A., Pretzer, G. M., & Warlaumont, A. S. (2020). Adult responses to infant prelinguistic vocalizations are associated with infant vocabulary: A home observation study. *PLOS ONE*, *15*(11), e0242232. https://doi.org/10.1371/journal.pone.0242232

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2022). *robustbase: Basic Robust Statistics* [Computer software]. http://robustbase.r-forge.r-project.org/

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

McGillion, M. L., Pine, J. M., Herbert, J. S., & 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), 1122–1131. https://doi.org/10.1111/jcpp.12725

Olson, J., & Masur, E. F. (2013). Mothers respond differently to infants' gestural versus nongestural communicative bids. *First Language*, *33*(4), 372–387. https://doi.org/10.1177/0142723713493346

Olson, J., & Masur, E. F. (2015). Mothers' labeling responses to infants' gestures predict vocabulary outcomes. *Journal of Child Language*, *42*(6), 1289–1311. https://doi.org/10.1017/S0305000914000828

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., van der Wal, J. J. M., Hulshoff Pol, H. E., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, *46*, 100868. https://doi.org/10.1016/j.dcn.2020.100868

Pearson, R. M., Heron, J., Melotti, R., Joinson, C., Stein, A., Ramchandani, P. G., & Evans, J. (2011). The association between observed non-verbal maternal responses at 12 months and later infant development at 18 months and IQ at 4 years: A longitudinal study. *Infant Behavior & Development*, *34*(4), 525–533. https://doi.org/10.1016/j.infbeh.2011.07.003

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur Communicative Development Inventory: Words and Sentences. *Journal of Child Language*, *27*(2), 255–266. https://doi.org/10.1017/S0305000900004098

Renzi, D. T., Romberg, A. R., Bolger, D. J., & Newman, R. S. (2017). Two minds are better than one: Cooperative communication as a new framework for understanding infant language learning. *Translational Issues in Psychological Science*, *3*, 19–33. https://doi.org/10.1037/tps0000088

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rowe, M. L., Ozçalişkan, S., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language*, *28*(2), 182–199. https://doi.org/10.1177/0142723707088310

Ruddy, M. G., & Bornstein, M. H. (1982). Cognitive correlates of infant attention and maternal stimulation over the first year of life. *Child Development*, *53*(1), 183–188. https://doi.org/10.2307/1129651

Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL*. Hartcourt Assessment BV.

Suarez-Rivera, C., Linn, E., & Tamis-LeMonda, C. S. (2022). From play to language: Infants' actions on objects cascade to word learning. *Language Learning*, *72*(4), 1092–1127. https://doi.org/10.1111/lang.12512

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science*, *23*(2), 121–126. https://doi.org/10.1177/0963721414522813

Verhagen, J., van den Berghe, R., Oudgenoeg-Paz, O., Küntay, A., & Leseman, P. (2019). Children's reliance on the non-verbal cues of a robot versus a human. *PLOS ONE*, *14*(12), e0217833. https://doi.org/10.1371/journal.pone.0217833

Wu, Z., & Gros-Louis, J. (2014). Infants' prelinguistic communicative acts and maternal responses: Relations to linguistic development. *First Language*, *34*(1), 72–90. https://doi.org/10.1177/0142723714521925

Wu, Z., & Gros-Louis, J. (2015). Caregivers provide more labeling responses to infants' pointing than to infants' object-directed vocalizations. *Journal of Child Language*, *42*(3), 538–561. https://doi.org/10.1017/S0305000914000221

Zink, I., & Lejaegere, M. (2002). *N-CDI's: Lijsten voor Communicatieve Ontwikkeling. Aanpassing en hernormering van de MacArthur CDI's van Fenson et al.* Acco.

Zink, I., & Lejaegere, M. (2003). *N-CDI's: Korte vormen, Aanpassing en hernormering van de MacArthur Short Form Vocabulary Checklist van Fenson et al.* Acco.

Chapter 6

General discussion and conclusions

The main goal of this dissertation was to predict variation in Dutch children's vocabulary skills by examining caregiver-infant interactions in a large, longitudinal cohort study. Previous studies have predominantly focused on the role of the caregiver in providing verbal responses to their infants' behaviours during caregiver-infant interactions, while we discussed in the general introduction (Chapter 1) that communication is **bidirectional** (i.e., involving an exchange of information between infant and caregiver) and **multimodal** (i.e., involving verbal and visual information). Visual cues in addition to verbal language might help infants to disambiguate speech and reduce referential ambiguity in the learning context – which could facilitate children's word learning. It remained largely unexplored whether verbal and multimodal dyadic behaviours are better predictors of children's concurrent and longitudinal vocabulary outcomes compared to children's individual behaviours. In this dissertation, we took a dyadic approach and studied infants' and caregivers' coupled verbal, nonverbal, and multimodal behaviours during caregiver-infant interactions – and their role in children's vocabulary outcomes. Given the large amount of manual annotation work that is required for analysing caregiver-infant interactions, we first explored to what extent we can use automated tools to facilitate the annotation process in Chapter 2. We also needed reliable measures of Dutch infants' and toddlers' vocabularies. In Chapter 3, we examined the reliability and validity of caregiver reports of Dutch children's vocabulary. Then, we examined whether demographic predictors of variation in children's vocabulary are age-specific or task-specific in this large, longitudinal cohort sample. In Chapter 4, we developed and tested a new coding scheme including caregivers' verbal and nonverbal responses, such as gestures and body orientation. Finally, we compared individual to dyadic behaviours to predict variation in Dutch children's concurrent and longitudinal vocabulary outcomes in Chapter 5. This last chapter summarises the main findings from the four empirical studies, discusses general implications, addresses methodological limitations, and proposes directions for future research.

6.1. Summaries of main findings

6.1.1. Overview of chapters

In Chapter 2, we aimed to examine the accuracy of open-source automatic speech recognition (ASR) tools for the annotation of Dutch infant-directed speech (IDS). There is a large interest in the annotation of speech addressed to infants. Therefore, we considered whether we can use open-source automated tools developed for adult-directed speech (ADS) to facilitate the manual annotation process of IDS. This could drastically speed up the time that is currently needed for creating manual annotations. However, the speech register IDS has specific acoustic properties, such as a higher mean pitch, a larger pitch range, and a slower speech rate, that might pose challenges for ASR tools developed for ADS. No previous studies had assessed the accuracy of such tools for the transcription of IDS. Since we needed an annotated corpus of Dutch IDS to assess the annotation accuracy, we used part of the cross-linguistic corpus of Dutch and Mandarin Chinese IDS for this study (Han, 2019). The first research question addressed in Chapter 2 was:

> **RQ 1: To what extent can we use open-source ASR tools to successfully transcribe Dutch IDS?**

To examine this, we first examined the accuracy of the open-source ASR tool Kaldi-NL at transcribing target words in IDS versus ADS. We found that Kaldi-NL correctly annotated only 55.8% of target words in IDS, while it annotated 66.8% correctly in ADS. We found significant negative effects of the speech register IDS and mean pitch on recognition accuracy. Given the low overall performance of Kaldi-NL on this dataset, we aimed to examine the difficulties in annotating IDS more broadly by comparing the word error rates (WERs) of full utterances, rather than target words alone, generated by two different open-source ASR tools: Kaldi-NL and WhisperX. When analysing full utterances, Kaldi-NL had a mean WER of 40.1%, while WhisperX had a mean WER of 22.5%. Depending on the goals of the research, these

automatic transcriptions should still be improved by a human annotator. However, the automatic transcriptions generated by WhisperX are of sufficient accuracy that they could be used to speed up the manual annotation process. This correction process will take less time compared to transcribing the data manually from scratch. While there is much room for improvement, automatic transcriptions generated by ASR tools developed for ADS therefore provide a promising start for researchers who have to transcribe large amounts of speech addressed to infants.

Before we can study predictors of variation in Dutch children's vocabulary skills, we need vocabulary measures that are reliable, valid, and sensitive enough to show variability across infants and toddlers. In Chapter 3, we first assessed the validity and reliability of the N$_{YOUth}$-CDIs of more than 300 Dutch children before we continued to use them to study individual differences in children's vocabulary outcomes. Then, we studied whether well-known demographic predictors of variation, such as maternal education and children's gender, are age-specific and task-specific in this large, longitudinal sample. Only a limited number of studies have examined the effects of these predictors longitudinally using multiple vocabulary measures across children's development, while there are questions regarding the onset and stability of these effects. The research questions were:

RQ 2: Are the N$_{YOUth}$-CDIs a valid and reliable measure of Dutch children's vocabulary?

RQ 3: Are demographic predictors of variation in children's vocabularies age-specific and task-specific?

Chapter 3 showed that the N$_{YOUth}$-CDIs show good reliability and strong concurrent and longitudinal validity. This suggests that we can reliably use the data to study variation in Dutch children's vocabulary size. The results indicated that for infants around 10 months of age, the gesture scale (i.e., indicating the size of the infant's

gesture and action repertoire) could be a more valid infant measure compared to the vocabulary scales (i.e., word production and word comprehension). Infants' gesture repertoire, as reported by their caregivers, was the only infant measure that was correlated with children's receptive vocabulary measured in the lab several years later. Then, we examined the effects of well-known demographic predictors of variation in children's vocabulary size. Although we found that children's gender, maternal education, and multilingualism explained some variance in children's vocabularies, the effects were age-specific and task-specific.

In Chapter 4, we presented a characterisation of 10-month-old infants' vocalisations and gestures and their caregivers' verbal, nonverbal, and multimodal (i.e., coordinated verbal and nonverbal) responses during six minutes of free play. Thus far, caregivers' nonverbal aspects of responsiveness remained largely understudied. We developed, trained, and tested a new coding scheme including caregivers' verbal, gestural, facial, and bodily responses. We also examined whether different infant behaviours tended to elicit different rates and types of caregivers' responses. The research questions driving this study were:

> **RQ 4: What types of caregivers' verbal, nonverbal, and multimodal responses to infants' vocalisations and gestures do we observe during free play?**
>
> **RQ 5: Do caregivers' verbal, nonverbal, and multimodal responses differ as a function of infants' vocalisations or gestures?**

First, we found that caregivers use a range of verbal, nonverbal, and multimodal behaviours when responding to their 10-month-old infants' vocalisations and gestures. Although the vast majority of responses were verbal (i.e., spoken language), approximately 40% of these verbal responses were multimodal – at least partially overlapping in time with a nonverbal behaviour. Then, we found that different infant

behaviours elicited different caregivers' response rates and types. The results showed that infants' bimodal behaviours (i.e., vocal-gestural combinations) elicited higher rates of verbal and multimodal responses from caregivers, while infants' gestures elicited higher rates of nonverbal responses. Overall, infants' vocalisations elicited the lowest response rates from caregivers. Furthermore, infants' index-finger points elicited more verbal responses from caregivers compared to other infants' gestures, while infants' gives elicited more nonverbal and multimodal responses from caregivers compared to other infant gestures. These findings suggest that infants play a role in shaping their learning environments by producing specific communicative behaviours, thereby influencing their caregivers' responsiveness.

Lastly, in Chapter 5, we assessed whether dyadic behaviours (i.e., combined infant behaviours and caregiver responses) were better predictors of children's vocabulary outcomes than infants' individual behaviours separately. To examine this, we contrasted the relative predictive value of three subsets of predictors on children's vocabulary outcomes: 1) infants' individual behaviours, 2) infants' behaviours met with caregivers' verbal responses, and 3) infants' behaviours met with caregivers' multimodal responses. We examined the behaviours around 9–11 months of age and used them to predict children's concurrent and longitudinal vocabulary outcomes measured around 2–4 years of age. No previous studies have directly contrasted different combinations of dyadic behaviours including different caregivers' response types. Therefore, it remained unexplored whether we find different effects when coupling infant behaviours with different types of caregiver responses. The research question driving the last study was:

> **RQ 6: Do dyadic combinations of infants' vocalisations and gestures (shows+gives and points) and caregivers' verbal and multimodal responses during free play improve the predictive value of infants' behaviours for children's vocabulary outcomes?**

We found that infants' points related to children's later receptive vocabularies, while infants' shows+gives related to children's later productive vocabularies – only when taking the instances that elicited caregivers' multimodal responses into account. We also found that only shows+gives which elicited verbal and multimodal responses from caregivers are positively related to infants' gesture repertoires. In conclusion, the dyadic combinations of infants' gestures and caregivers' responses have more predictive value for children's concurrent and longitudinal vocabulary outcomes than infants' individual behaviours separately. This highlights the importance of studying dyadic behaviours during caregiver-infant interactions.

6.2. General discussion, implications, and future research

The gaps in the literature concerned methodological and theoretical issues related to data annotation, vocabulary measures, the onset and stability of predictors of vocabulary size, and the unknown influences of dyadic and multimodal behaviours during caregiver-infant interactions on children's vocabulary outcomes. The next section discusses the broader implications of the research findings presented in this dissertation and possible directions for future studies.

6.2.1. Automated tools can facilitate data annotation

Tools are being developed to generate automatic annotations that would greatly benefit research on IDS by speeding up the annotation process (Burnham et al., 2016). Typically, research with infants is limited to small samples because it is difficult, expensive, and time-consuming to recruit infants (Oakes, 2017). Longitudinal cohort studies provide excellent opportunities for researchers to include larger sample sizes in their studies. Overall, larger sample sizes give more statistical power which results in smaller margins of error and more reliable findings. Besides recruiting infants and collecting data, however, researchers still need to annotate the large amounts of raw data. This is an expensive and time-consuming task. In studies examining speech data or caregiver-infant interactions, such as the ones reported in this dissertation, researchers typically have to annotate large amounts of audio and/or video data. Segmenting, annotating, and transcribing an hour of speech can take up to fifty hours

in total (Barras et al., 2001). Annotating videos frame by frame may take even longer, depending on the number of different behaviours being analysed. Therefore, exploring whether we can use automated tools to facilitate the annotation process of IDS is of great interest to researchers studying language input. The results of Chapter 2 suggest that we can use automated tools to facilitate the annotation process. Although the features of IDS improve word recognition abilities in infants (e.g., Estes & Hurley, 2013; Song et al., 2010), we found that IDS decreases word recognition accuracy in ASR tools developed for ADS. Yet, some state-of-the-art tools trained on large amounts of semi-supervised training data (e.g., WhisperX) are sufficiently accurate to support the transcription of Dutch IDS.

These results have several implications. First, we would advise researchers to try several different open-source ASR tools that are available in the target language because the results in Chapter 2 suggest that different ASR tools can generate automatic transcriptions of vastly different accuracy levels (Kaldi-NL versus WhisperX). WhisperX is trained on a much larger semi-supervised cross-linguistic data set compared to Kaldi-NL which could be better equipped to encompass the large acoustic variability found in IDS. Second, we would advise researchers to use automated tools as a starting point for the manual annotation process. The automatic transcriptions might not be sufficiently accurate for all research purposes. However, researchers can still save time by manually correcting the automatic transcriptions rather than transcribing all speech data manually from scratch. Lastly, we would encourage researchers to improve the recognition accuracy of ASR tools for the annotation of IDS in future studies. Since the results of our study suggest that a higher mean pitch results in more word recognition errors, researchers could try to apply front-end lowering of the mean pitch of IDS recordings to improve the recognition accuracy (see Gustafson & Sjölander, 2002 for application of this method to children's speech). Another approach could be to train completely new language and/or acoustic models on IDS corpora. Previous studies show that matching training and testing data results in the highest performance accuracy (Kirchhoff & Schimmel, 2005). For this

to work, the IDS corpora must be large and general enough to be useful for application on new datasets.

While we can use WhisperX to facilitate the manual annotation process of Dutch IDS, we did not have the chance to apply this method when annotating the caregiver-infant interactions used in this dissertation. We are currently working with student assistants to manually improve the automatic transcriptions generated by WhisperX of the caregiver-infant interactions collected in the YOUth study. This will allow us to determine whether the promising results generalise to other data sets.

6.2.2. Reports of infants' gestures predict later vocabulary

We established the reliability and concurrent and longitudinal validity of the NYOUth-CDIs. We found that for infants around 9–11 months, only the gesture scale correlated significantly with children's receptive vocabulary scores measured in the lab around 2–4 years of age. Infants' word production and word comprehension positively correlated with children's word production at 2–4 years as reported by their caregivers, but not with the lab-administered task. The finding that infants' gesture repertoires at 9–11 months as reported by their caregivers significantly correlates with a lab-administered task is strong evidence for its predictive validity. In addition, we found that maternal education only negatively affected infants' word comprehension and word production, but not infants' gesture repertoires. The negative effects of maternal education on the vocabulary scales for infants could be caused by reporting biases. Determining whether an infant can make certain gestures requires less interpretation for caregivers than determining whether an infant can comprehend or produce words, especially because CDIs also accept mispronunciations. Caregivers of lower socio-economic status could overreport their children's vocabulary skills when they think larger vocabularies are more desirable or when their criteria for word comprehension and production are more liberal than others', for example. The absence of a negative effect of maternal education on infants' gesture repertoires suggests that this scale is less affected by reporting biases – possibly because gestures are more discrete, easily observable, and less affected by societal expectations.

However, future studies with more diverse samples should carefully examine our assumption that the negative effects of maternal education on infants' word production and word comprehension are only caused by reporting biases. The question of whether there are already facilitative effects of maternal education on infants' vocabulary skills, or whether any effects only emerge later in development, remains unanswered (for a discussion, see section 6.2.3 below). In sum, the gesture scale seems to be a valid method for capturing individual differences across infants aged 9–11 months which can predict children's vocabulary skills measured years later. We would therefore recommend researchers who aim to capture variability in infants' vocabulary skills to also include the gesture scale when administering CDIs. The gesture scale is not normally included in CDI short forms.

The gesture scale has stronger predictive value for children's later receptive vocabulary compared to the vocabulary scales for infants. The gesture scale may be the only CDI scale that shows enough variability across infants aged 9–11 months – making it possible to detect any small effects, such as an advantage for girls. Besides methodological implications, this result also has a theoretical implication. Previous studies have often found that differences across infants' gestures explain differences in their later vocabularies (e.g., Brooks & Meltzoff, 2008; Colonnesi et al., 2010; Rowe & Goldin-Meadow, 2009). The positive relationship between infants' gesture repertoires and their later receptive vocabulary skills found in Chapter 3 could be driven by the presence of specific gestures in infants' gesture repertoires, for example, index-finger pointing. Infants' index-finger pointing is a robust predictor of their vocabulary outcomes as reported in Chapter 5 (also see Colonnesi et al., 2010). Future studies should examine whether the sheer sizes of infants' gesture repertoires can predict children's later vocabularies or whether this effect is driven by the presence of specific types of gestures in infants' gesture repertoires. Nevertheless, the results imply that caregiver reports of infants' gestures are predictive of their later vocabulary outcomes.

6.2.3. Examine multiple vocabulary measures over time

We used this large longitudinal sample to study the effects of well-known demographic predictors across different developmental stages and vocabulary outcome measures. Limited studies have examined demographic differences in vocabulary over time, while there are questions regarding the onset and stability of the effects. For example, studies report positive effects of maternal education on toddlers' vocabularies (Feldman et al., 2000; Fenson et al., 2007; but cf. Reese & Read, 2000), while studies often report negative effects of maternal education on infants' vocabularies (Bavin et al., 2008; Feldman et al., 2000; Reese & Read, 2000). The effects of demographic predictors of variation in children's vocabularies vary based on differences in sample characteristics and/or the vocabulary measure being used. This is the reason why we examined these predictors within one longitudinal sample for which we collected multiple different vocabulary outcome measures during infancy and toddlerhood provides us with an excellent opportunity to examine whether such effects are age-specific or task-specific while keeping the sample constant.

The results of Chapter 3 suggest that widely reported demographic predictors of variation in children's vocabulary outcomes are age-specific and task-specific. The strengths and directions of the effects of predictors vary across children's developmental stages. Such developmental changes were often found in previous longitudinal studies. For example, 10-month-old infants' show+gives were better predictors of later language skills, while four months later, points were better predictors (Choi et al., 2021). In addition, caregivers' quantity of input is more important during the second year of life, while caregivers' quality (e.g., diversity or decontextualised utterances) of input is more important during the third year of life (Rowe, 2012). Moreover, the facilitating effects of maternal education on children's vocabulary skills may only emerge later in development (Rowland et al., 2022). The broader implication is that, when examining factors that influence children's vocabulary, it can be more insightful to examine children's vocabulary outcomes

longitudinally across development. The influencing factors could change or emerge only later in development.

The results of Chapter 3 also suggest that the predictors may vary depending on the type of vocabulary outcome (i.e., word production, word comprehension, or gesture repertoire) or measurement method (i.e., caregiver reports versus lab-administered tasks). This has two possible explanations. The first explanation is that factors explaining variation in children's vocabularies may have differential effects on the sizes of children's expressive vocabulary, comprehensive vocabulary, and gesture repertoires. For example, the advantage for girls compared to boys seems to be more prominent in word production than word comprehension (also see Frank et al., 2021). In Chapter 3, we showed that an early advantage for girls seems to appear in infants' gesture repertoires, while at this early age, girls and boys do not yet differ in their word production or word comprehension skills. The gesture scale may be a more suitable measurement for infants at this early age due to floor effects in the vocabulary scales, as discussed above. It is also possible that the gender effect starts as a difference in infants' gestures, subsequently influencing children's later vocabulary outcomes. Nevertheless, some factors likely influence certain aspects of vocabulary more than others. The second explanation is that all vocabulary measures collected during infancy were caregiver-reported measures which can be influenced by caregiver reporting biases. For example, we found negative effects of maternal education on infants' word production and word comprehension. We did not find an effect of maternal education on toddlers' word production as reported by their caregivers. Possibly, reports of infants' knowledge require more interpretation than reports of toddlers' knowledge. We found a positive effect of maternal education on the lab-administered task which cannot be influenced by reporting biases. The broader implication of these findings is that, when examining the effects of a factor explaining variation in children's vocabularies, researchers can increase the validity of their findings by including multiple vocabulary outcome measures over time. We would particularly encourage the use of a lab-administered measure of vocabulary alongside caregiver reports.

6.2.4. Infant gestures which elicited responses predict later vocabulary

Previous studies have found that infants' prelinguistic behaviours, including shows, gives, and points, are predictors of their vocabulary outcomes (Choi et al., 2021; Colonnesi et al., 2010). Infants' early deictic gestures could predict their vocabulary outcomes because they teach infants about the connection between a symbol and its referent, they establish joint attention, and/or they elicit contingent responses from caregivers (e.g., Bruner, 1983; Choi et al., 2021; Colonnesi et al., 2010; Donnellan et al., 2019). In line with the latter hypothesis, studies have found that caregivers individually differ in their verbal responsiveness to infants' prelinguistic behaviours which has been found to positively predict children's vocabulary outcomes (e.g., Donnellan et al., 2019; Olson & Masur, 2015; Wu & Gros-Louis, 2014). It has been suggested that caregivers' contingent verbal labelling responses mediate the relationship between infants' gestures and children's vocabulary outcomes (Olson & Masur, 2015). There is robust evidence that individual behaviours (infants' prelinguistic behaviours and caregivers' responses) separately predict children's later vocabulary, while it remained understudied how caregivers and infants jointly contribute to create the word-learning context. In Chapter 4, we found that infants' points tend to elicit more verbal responses from caregivers (also see Wu & Gros-Louis, 2015) which could be the mechanism through which these gestures are positively related to children's later vocabulary outcomes (Olson & Masur, 2015). This hypothesis is further supported by the finding that infants' index-finger points during caregiver-infant interactions were related to children's later receptive vocabularies, as reported in Chapter 5. We also found that infants' gives tend to elicit higher rates of nonverbal and multimodal responses (i.e., verbal responses at least partially overlapping with a nonverbal behaviour) in Chapter 4. If 10-month-old infants shows+gives are predictors of their later vocabularies because they tend to elicit more responses from caregivers at this age, as hypothesised by Choi et al. (2021), then we would expect that caregivers' multimodal responses facilitate children's vocabulary learning. We found support for this hypothesis. In Chapter 5, we showed that the dyadic combination of infants' shows+gives and caregivers' multimodal responses are related to children's later productive vocabularies, while the

total of infants' shows+gives regardless of caregivers' responses were not. This finding highlights the importance of studying dyadic behaviours during caregiver-infant interactions.

Caregivers' verbal, nonverbal, and multimodal responses vary as a function of infants' prelinguistic behaviours as shown in Chapter 4. This could explain why different gestures differentially relate to children's vocabulary outcomes. For example, in Chapter 5, we showed that infants' points, which frequently elicit verbal responses from caregivers, can predict children's word comprehension skills, while infants' shows+gives, which frequently elicit caregivers' multimodal responses, can predict children's word production skills. When infants repeatedly receive verbal object labels from caregivers after using their index-finger to point towards objects, they learn about the referential function of points. The comprehension of points could help infants to reduce referential ambiguities in the learning context. When caregivers hand over objects after the infant has reached for them while opening and closing their fingers, the infant learns the communicative intent of a requesting gesture. The reciprocal nature of the infant requesting, the caregiver passing, and the infant then accepting the toy could tap into children's expressive language development. Research has shown that infants' conversational turn-taking influences their expressive vocabulary development (e.g., Donnelly & Kidd, 2021). When infants actively start more interactions, caregivers have more opportunities to scaffold language by engaging with the objects in a feedback loop (Bruner, 1983; Tamis-LeMonda et al., 2014). This could explain why infant gestures differentially predict children's vocabulary development. Because caregivers tend to respond in a timely and appropriate manner, children learn about the reciprocal nature of communication and the meanings of different gestures. This could facilitate children's vocabulary development because infants learn how to retrieve the information that they want (i.e., improve their information-seeking behaviours) and caregivers have more opportunities to reply appropriately (i.e., improve their information-providing behaviours) which creates clearer signals for word learning (also see Chen et al., 2021).

The broader implication of these results is that when analysing caregiver-infant interactions, it is more informative to study the behaviours of both interlocutors. When analysing caregiver responsiveness, previous studies typically distinguish between high and low levels of caregiver responsiveness. This is then treated as a characteristic of the caregiver. As shown by the results in this dissertation, however, these findings can be strongly influenced by the infant. If some infants produce fewer gestures than others, they will elicit fewer responses from their caregivers, as shown in Chapter 4. Therefore, examining dyadic behaviours during interactions results in a more complete picture. The next step is to research how caregivers' responses in turn influence the production of infants' prelinguistic behaviours. As previously shown by Ger et al. (2018), caregivers' contingent responses to infants' pointing at 10 months predicted an increase in infants' pointing at 12 months. By reinforcing infants' behaviours over time, this presupposes that infants in turn elicit more responses from their caregivers. This could create a feedback loop that is efficient for learning. Similar results concerning the importance of the bidirectional nature of caregiver-infant interactions have been found for infants' vocal development. Goldstein and Schwade (2008) have shown that infants use caregivers' contingent verbal responses to their babbles to restructure their vocalisations to match the phonological patterns that they heard in their caregivers' speech. The bidirectional nature of interactions underscores the importance of studying coupled dyadic behaviours during social interactions to better understand the early learning environment.

6.2.5. Caregivers' multimodal responses influence vocabulary outcomes

Previous studies have largely focused on the role of the caregiver in providing verbal responses to infant behaviours, while other strands of research emphasise the importance of caregivers' multimodal communication. For example, children use nonverbal cues, such as pointing gestures, that co-occur with caregivers' speech to learn the reference of novel or unclear speech (e.g., Baldwin et al., 1996; Grassmann & Tomasello, 2010). Recently, Chen et al. (2021) showed that caregivers touched objects more often while naming them only when the objects were unfamiliar to the child. In Chapter 4, we showed that 40% of caregivers' verbal responses to infants'

behaviours were multimodal (i.e., verbal + nonverbal). We found particularly many verbal responses accompanied by manual gestures. To a lesser extent, we found that caregivers also coordinate speech with facial expressions or other bodily behaviours, such as repositioning themselves to face the toy of interest to the child. Our study extends previous findings by showing that we find the same degree of multimodality in caregivers' responsiveness as found in caregivers' overall communication (e.g., Vigliocco et al., 2019). To what extent do caregivers' multimodal responses contribute to children's vocabulary development in addition to their verbal responses?

In this dissertation, we examined the influence of caregivers' verbal and multimodal responses to infants' points, shows+gives, and vocalisations on children's vocabulary outcomes. In Chapter 4, we found that infants' gives elicited more multimodal responses from caregivers compared to other infant gestures. In Chapter 5, we showed that only infants' shows+gives which elicited caregivers' multimodal responses – but not verbal responses – are positively related to children's later productive vocabulary. This finding suggests that it is not exclusively caregivers' verbal responsiveness that contributes to children's vocabulary development, but possibly all types of responsiveness if it is appropriate to the infant gesture. Caregivers' verbal and multimodal responses could differentially relate to children's vocabulary outcomes. Caregivers' verbal responsiveness seems more strongly related to children's language comprehension, while caregivers' multimodal responsiveness could be more strongly related to children's language production. The results presented in Chapter 5 align with this hypothesis. This is consistent with the findings by Choi et al. (2021) for shows+gives. In their study, they found that 10-month-old infants' shows+gives only correlated with their later expressive – but not receptive – vocabulary outcomes. The reciprocal and multimodal nature of infants' giving a toy, caregivers' accepting the toy, playing with it and commenting on it, and handing the toy back to the infant could tap into infants' conversational turn-taking skills which influence their expressive vocabulary outcomes (Donnelly & Kidd, 2021). In contrast, infants' points tend to elicit verbal responses. Studies typically find that infants' pointing gestures are related to both vocabulary production and comprehension, although meta-analyses indicate

that the relationship might be stronger with children's receptive vocabulary outcomes (Colonnesi et al., 2010; Kirk et al., 2022). It could be possible that infants are more likely to develop language comprehension skills from caregivers' contingent verbal responses, while they develop expressive language abilities from caregivers' interactive, multimodal behaviours that are highly present in give-and-take sequences during free play.

The broader implication of these findings is that caregivers' multimodal language input could play a unique role in facilitating children's language development. The high prevalence of multimodal communication in caregiver-infant interactions highlights the need to annotate both the verbal and nonverbal modalities when examining the effects of caregivers' language input on children's vocabulary development. When examining only the verbal domain, researchers miss relevant information shaping children's early learning environments. Future studies can further examine which nonverbal behaviours produced by caregivers are relevant to children's vocabulary outcomes. These are likely to include behaviours which contain useful referential information, such as index-finger pointing. The combination of speech and a visual referential cue might make it easier for the infant to map the phonological form of novel words onto their referents, subsequently facilitating word learning.

6.2.6. Summary of new insights

In sum, this dissertation presents new insights into data annotation, vocabulary measurements, and the role of dyadic and multimodal aspects of caregiver-infant interactions in predicting children's vocabulary outcomes. For the first time, we showed that researchers can make use of existing automated tools to facilitate the manual annotation process of speech addressed to infants. This can drastically speed up research in this area. Concerning vocabulary measurements, we provided evidence for the long-term predictive validity of caregiver reports of infants' vocabulary in a large, longitudinal cohort sample. Already at 9–11 months of age, caregiver reports of infants' gesture repertoires are sensitive and reliable enough to predict children's

receptive vocabulary outcomes measured in the lab years later. The last two empirical chapters were the first studies to extensively document caregivers' verbal, nonverbal and multimodal responses to infants' vocalisations and gestures. While infants' deictic gestures predict their vocabulary outcomes, we found that caregivers' verbal and multimodal responses can improve the predictive relationships. This was the first study to show that caregivers' multimodal responsiveness contributes to children's expressive vocabulary development. Another significant contribution is that this dissertation provided evidence for the long-term predictive validity of these measures. Infants' gestures and caregivers' responses during infancy remain predictive of children's vocabulary outcomes years later. The results reveal the importance of studying dyadic and multimodal behaviours during caregiver-infant interactions when predicting children's long-term vocabulary outcomes.

6.3. Methodological limitations

Despite the advantages that emerged from using data collected within a large, longitudinal cohort study, there are four limitations. The first limitation concerns the homogeneous sample with an overrepresentation of highly educated families – which is a common problem. Lower educated families tend to participate less frequently in scientific studies. This might raise questions about the generalisability of the findings to lower educated samples. For example, in Chapter 3, we analysed the effects of maternal education on children's longitudinal vocabulary outcomes. Although our sample presented a good opportunity to study the effects of maternal education in a large sample with multiple longitudinal vocabulary outcomes, which allowed us to examine the stability of the effect over time, the sample may not have been diverse enough to show the effects of maternal education in all outcome measures. We did not find any SES differences in infants' gesture repertoires in Chapter 3, while we did find negative effects of maternal education and infants' pointing frequency on infants' word comprehension skills in Chapter 5. In Chapter 5, we hypothesised that the negative SES-effect on infants' word comprehension is likely caused by a caregiver reporting bias. Based on the literature, we would expect children of higher SES families to produce more gestures (e.g., Ger et al., 2023; Rowe & Goldin-Meadow,

2009). This could suggest that the negative effect that we found for infants' pointing frequency is moderated by the negative SES-effect. The low frequency of infants' pointing gestures in our data, in combination with the low prevalence of lower SES families in our sample, does not allow us to reliably examine whether children of higher SES families produced more pointing gestures nor test for an interaction between the negative effects of maternal education and infants' pointing frequency on infants' word comprehension skills. In future studies with more representative samples, researchers could examine whether children from higher SES families tend to produce more pointing gestures compared to children from lower SES families (as recently reported by Ger et al., 2023), and possibly whether this explains the negative relationship between the frequency of infants' pointing gestures and their word comprehension skills as reported by their caregivers.

The second limitation concerns the durations of observations. There were only six minutes of free play recorded per dyad during the caregiver-infant observations in the lab. This is a short time that may not fully represent the child's abilities (and subsequently caregivers' responsiveness). A recent meta-analysis on the effects of caregivers' input quality and quantity on children's language skills suggests that effect sizes increase when the observation length is longer (Anderson et al., 2021; cf. Madigan et al., 2019 on parenting behaviours). Longer observations are more representative. Nevertheless, we still found significant effects of infants' pointing frequency and infants' shows+gives combined with caregivers' multimodal responses at 9–11 months on children's long-term vocabulary outcomes at 2–4 years in Chapter 5. This likely indicates that the effects of these individual and dyadic behaviours on children's vocabulary skills are robust. We did not find any evidence that infants' prelinguistic vocalisations, combined with caregivers' contingent responses or not, are related to children's vocabulary skills. This does not align with earlier findings (Donnellan et al., 2019; Lopez et al., 2020). Infants' prelinguistic vocalisations may be weaker predictors of children's vocabulary outcomes compared to infants' gestures, and the observation period in our study could have been too short to reliably detect small effects. Then, the observations were also recorded in a lab setting.

Although the caregivers were instructed to play with their infants as if they were at home, resulting in semi-naturalistic play, the data are not as representative as they would have been when they were collected in the home environment (see Lopez et al., 2020).

The third limitation concerns the developmental timepoint during which we analysed the caregiver-infant interactions. First, we only analysed caregiver-infant interactions when the infants were 9–11 months of age. Children generally start producing their first gestures around 9 months of age (Frank et al., 2021). Therefore, the frequency and variation of infants' gestures were limited in our data set. When designing our studies, we were limited by the methodological choices made in the YOUth study. For our purposes, more caregiver-infant recordings and vocabulary measures obtained during the second year of life – a crucial year of rapid vocabulary development – would have been more appropriate. Yet, this also resulted in a strong point: The results in Chapter 5 show that infants' earliest shows+gives and points, measured at 9–11 months, already predict their long-term vocabulary outcomes at 2–4 years of age. This suggests there is a strong predictive relationship. Second, we only analysed caregiver-infant interactions at one point in time. A recent study showed that at 10 months of age, infants' shows+gives were better predictors of children's later vocabulary outcomes than points. By 14 months, however, infants' points were better predictors (Choi et al., 2021). The authors also found that at 10 months, caregivers responded more often to shows+gives than points, while this difference had disappeared by 14 months. The best dyadic combinations of infants' behaviours and caregivers' responses to predict children's vocabulary outcomes may change across children's development. Therefore, future studies should examine different dyadic combinations across children's development and evaluate their predictive value for children's vocabulary outcomes.

Lastly, due to time constraints we were limited by the number of social cues that we could annotate during the free play sessions. It could have been informative to also annotate eye gaze. There is evidence that infants are more likely to learn word-object

relations in the presence of eye gaze cues (see Çetinçelik et al., 2021). In addition, gaze checking (i.e., looking into someone's eyes) or gaze alternation (i.e., changing gaze between the caregiver and the object of interest) while gesturing could mark communicative intent during interactions (e.g., Bates et al., 1975; Tomasello et al., 1997; Wu & Gros-Louis, 2014). A study found that infants' gaze-coordinated behaviours are more likely to elicit responses from caregivers, which in turn are more likely to facilitate children's vocabulary outcomes (Donnellan et al., 2019). Yet, previous studies also note that infants tend to pay closer attention to spatially precise manual cues, including manual actions with objects, rather than eye gaze (Verhagen et al., 2019; Yu & Smith, 2013) and eye contact or attention to the speaker does not facilitate infants' neural tracking of speech either (Çetinçelik et al., 2023). Nevertheless, caregivers' contingent verbal responses coordinated with eye gaze to the object of interest (or gaze checking between the infant and the object) could further improve children's learning outcomes.

6.4. Conclusions

The overarching goal of this dissertation was to predict variation in Dutch children's vocabulary skills by examining caregiver-infant interactions in a large, longitudinal cohort study. We addressed research gaps concerning data annotation, vocabulary measurements, and the role of dyadic and multimodal aspects of caregiver-infant interactions in predicting children's vocabulary. First, while currently there is a large interest in the annotation of speech addressed to infants, the accuracy of existing automated tools for the annotation of IDS has thus far remained unexplored. We showed that researchers can successfully use existing tools to facilitate the manual annotation process. Second, after establishing the validity and reliability of the NYOUth-CDIs, we showed that well-known demographic factors influencing variation in children's vocabularies, such as maternal education or children's gender, are age-specific and task-specific. This highlights the importance of including multiple vocabulary outcome measures across children's development. Third, the results highlight the importance of studying dyadic behaviours when examining caregiver-infant interactions. During interactions, infants and caregivers both shape and are

shaped by each other's behaviours. The results of our study suggest that the dyadic combinations of infants' behaviours and caregivers' responses are better predictors of children's vocabulary outcomes compared to infants' individual behaviours. Lastly, we show the importance of studying caregivers' nonverbal, in addition to verbal, behaviours during caregiver-infant interactions. Caregivers' multimodal responses to infants' gestures could play a unique role in children's expressive vocabulary development. In research on children's vocabulary development, we aim to describe how infants gather sufficient information from the language input that allows them to learn words. Studying the dyadic and multimodal nature of early caregiver-infant interactions creates a more complete picture of children's learning environments which brings us closer to solving this puzzle.

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*(2), 484–501. https://doi.org/10.1111/cdev.13508

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, *67*(6), 3135–3153. https://doi.org/10.2307/1131771

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5–22. https://doi.org/10.1016/S0167-6393(00)00067-4

Bates, E., Camaioni, L., & Volterra, V. (1975). The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly of Behavior and Development*, *21*(3), 205–226.

Bavin, E. L., Prior, M., Reilly, S., Bretherton, L., Williams, J., Eadie, P., Barrett, Y., & Ukoumunne, O. C. (2008). The Early Language in Victoria Study: Predicting vocabulary at age one and two years from gesture and object use. *Journal of Child Language*, *35*(3), 687–701. https://doi.org/10.1017/S0305000908008726

Bornstein, M. H., Tamis-Lemonda, C. S., Hahn, C.-S., & Haynes, O. M. (2008). Maternal responsiveness to young children at three ages: Longitudinal analysis of a multidimensional, modular, and specific parenting construct. *Developmental Psychology*, *44*(3), 867–874. https://doi.org/10.1037/0012-1649.44.3.867

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2019). Intention or attention before pointing: Do infants' early holdout gestures reflect evidence of a declarative motive? *Infancy*, *24*(2), 228–248. https://doi.org/10.1111/infa.12267

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/s030500090700829x

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

Burnham, D., Kalashnikova, M., Muawiyath, S., Cassidy, S., & Estival, D. (2016). *Infant-directed speech research made easy: A database, some tools and a virtual laboratory*. 43rd Experimental Psychology Conference, Melbourne, Australia.

Çetinçelik, M., Rowland, C. F., & Snijders, T. M. (2021). Do the eyes have it? A systematic review on the role of eye gaze in infant language development. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.589096

Çetinçelik, M., Rowland, C. F., & Snijders, T. M. (2023). Ten-month-old infants' neural tracking of naturalistic speech is not facilitated by the speaker's eye gaze. *Developmental Cognitive Neuroscience*, *64*, 101297. https://doi.org/10.1016/j.dcn.2023.101297

Chen, C., Houston, D. M., & Yu, C. (2021). Parent–child joint behaviors in novel object play create high-quality data for word learning. *Child Development*, *92*(5), 1889–1905. https://doi.org/10.1111/cdev.13620

Choi, B., Wei, R., & Rowe, M. L. (2021). Show, give, and point gestures across infancy differentially predict language development. *Developmental Psychology*, *57*(6), 851–862. https://doi.org/10.1037/dev0001195

Colonnesi, C., Stams, G. J. J. M., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366. https://doi.org/10.1016/j.dr.2010.10.001

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2019). Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants' vocalizations, gestures and word production. *Developmental Science*, *23*(1), 1–21. https://doi.org/10.1111/desc.12843

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, *92*(2), 609–625. https://doi.org/10.1111/cdev.13511

Estes, K. G., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*(2), 310–322. https://doi.org/10.1111/1467-8624.00146

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur Communicative Development Inventories: User's guide and technical manual* (Second edition). Paul H. Brookes Publishing Co.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank Project*. MIT Press. https://langcog.github.io/wordbank-book/

Ger, E., Altınok, N., Liszkowski, U., & Küntay, A. C. (2018). Development of infant pointing from 10 to 12 months: The role of relevant caregiver responsiveness. *Infancy*, *23*(5), 708–729. https://doi.org/10.1111/infa.12239

Ger, E., Küntay, A. C., Ertaş, S., Koşkulu-Sancar, S., & Liszkowski, U. (2023). Correlates of infant pointing frequency in the first year. *Infancy*, 1–21. https://doi.org/10.1111/infa.12560

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*(4), 878–894. https://doi.org/10.1111/1467-8624.00197

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science*, *13*(1), 252–263. https://doi.org/10.1111/j.1467-7687.2009.00871.x

Gustafson, J., & Sjölander, K. (2002). Voice transformations for improving children's speech recognition in a publicly available dialogue system. *7th International Conference on Spoken Language Processing*, 297–300. https://doi.org/10.21437/ICSLP.2002-139

Han, M. (2019). *The role of prosodic input in word learning: A cross-linguistic investigation of Dutch and Mandarin Chinese infant-directed speech* [Dissertation, Utrecht University]. http://localhost/handle/1874/379614

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4 Pt 1), 2238–2246. https://doi.org/10.1121/1.1869172

Kirk, E., Donnelly, S., Furman, R., Warmington, M., Glanville, J., & Eggleston, A. (2022). The relationship between infant pointing and language development: A meta-analytic review. *Developmental Review*, *64*, 101023. https://doi.org/10.1016/j.dr.2022.101023

Lopez, L. D., Walle, E. A., Pretzer, G. M., & Warlaumont, A. S. (2020). Adult responses to infant prelinguistic vocalizations are associated with infant vocabulary: A home observation study. *PLOS ONE*, *15*(11), e0242232. https://doi.org/10.1371/journal.pone.0242232

Madigan, S., Prime, H., Graham, S. A., Rodrigues, M., Anderson, N., Khoury, J., & Jenkins, J. M. (2019). Parenting behavior and child language: A meta-analysis. *Pediatrics*, *144*(4). https://doi.org/10.1542/peds.2018-3556

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

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186

Olson, J., & Masur, E. F. (2015). Mothers' labeling responses to infants' gestures predict vocabulary outcomes. *Journal of Child Language*, *42*(6), 1289–1311. https://doi.org/10.1017/S0305000914000828

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur Communicative Development Inventory: Words and Sentences. *Journal of Child Language*, *27*(2), 255–266. https://doi.org/10.1017/S0305000900004098

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., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rowland, C., Krajewski, G., Meints, K., Łuniewska, M., Kochańska, M. K., & Alcock, K. (2022). *CDI Demographics*. https://osf.io/hwg4c/

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

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science*, *23*(2), 121–126. https://doi.org/10.1177/0963721414522813

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, *68*(6), 1067–1080.

Verhagen, J., van den Berghe, R., Oudgenoeg-Paz, O., Küntay, A., & Leseman, P. (2019). Children's reliance on the non-verbal cues of a robot versus a human. *PLOS ONE*, *14*(12), e0217833. https://doi.org/10.1371/journal.pone.0217833

Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C. R., Milan Maillo, I., & Perniss, P. (2019). Onomatopoeias, gestures, actions and words in the input to children: How do caregivers use multimodal cues in their communication to children? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 7.

Wu, Z., & Gros-Louis, J. (2014). Infants' prelinguistic communicative acts and maternal responses: Relations to linguistic development. *First Language*, *34*(1), 72–90. https://doi.org/10.1177/0142723714521925

Wu, Z., & Gros-Louis, J. (2015). Caregivers provide more labeling responses to infants' pointing than to infants' object-directed vocalizations. *Journal of Child Language*, *42*(3), 538–561. https://doi.org/10.1017/S0305000914000221

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLOS ONE*, *8*(11), e79659. https://doi.org/10.1371/journal.pone.0079659

Nederlandse samenvatting
(Summary in Dutch)

Ouder-kind interacties en de woordenschat van kinderen: Een grootschalig, longitudinaal onderzoek naar dyadische en multimodale gedragingen

Kinderen worden geboren in een complexe wereld vol auditieve en visuele signalen waaruit ze regels en betekenissen moeten halen. Om effectief te kunnen communiceren, moeten kinderen de namen van objecten, acties en gebeurtenissen leren. Na het segmenteren van woorden uit het continue spraaksignaal, moeten kinderen de betekenissen van deze woorden leren. Dit is een computationeel probleem: Als kinderen een nieuw woord horen, zien ze oneindig veel mogelijke referenten in hun omgeving. Kinderen moeten deze referentiële ambiguïteiten oplossen voor het leren van hun moedertaal. Voorgaande studies hebben onderzocht welke informatie kinderen kunnen gebruiken om dit voor elkaar te krijgen. Kinderen kunnen bijvoorbeeld sociale signalen gebruiken, zoals aanwijsgebaren of lichaamsoriëntatie (e.g., Baldwin et al., 1996; Grassmann & Tomasello, 2010). Wanneer de ouder een nieuw woord zegt en tegelijkertijd naar de referent wijst heeft het kind meer aanwijzingen waar het nieuwe woord naar refereert. Sociale interacties zijn rijk aan zulke signalen die kinderen kunnen helpen met het leren van taal.

Bestaande theorieën benadrukken het belang van sociale interacties voor het leren van taal (Bruner, 1983; Vygotsky, 1962). Veel eerdere onderzoeken laten de impact van sociale factoren op de taalontwikkeling van kinderen zien (zie reviews: Hoff, 2006; Kuhl, 2007; Rowe & Weisleder, 2020). Sociale interacties worden gekenmerkt door contingentie. Als een kind bijvoorbeeld naar een pop reikt en de ouder/verzorger deze overhandigt en zegt: “Hier is de pop”, dan is de reactie van de ouder/verzorger contingent omdat deze snel en passend is bij de actie van het kind (Skinner, 1986;

Tamis-LeMonda et al., 2014). Eerdere studies hebben vier kenmerken geopperd waardoor contingente reacties van ouders/verzorgers het leren van taal kunnen vergemakkelijken: de reacties zijn temporeel contingent, semantisch contingent, pragmatisch contingent en aandachtsgericht (Kuhl, 2007; Masek et al., 2021a; Tamis-LeMonda et al., 2014). Ten eerste maakt de temporele contingentie van de reacties van de ouder/verzorger op de actie van een kind – wanneer de reactie van de ouder kort na het gedrag van het kind plaatsvindt – het voor het kind gemakkelijker om te begrijpen dat de twee gebeurtenissen direct aan elkaar zijn gelinkt (Jaffe et al., 2001; Keller et al., 1999). Ten tweede zou de semantische contingentie – wanneer de inhoud van de reactie gerelateerd is aan het object of de activiteit waar het kind mee bezig is – het voor het kind gemakkelijker kunnen maken om de woorden te koppelen aan de omgeving (Baldwin & Markman, 1989; Carpenter et al., 1998). Ten derde kunnen contingente reacties een pragmatische functie hebben. Baby's leren dat ze effectief kunnen communiceren door vocalisaties en gebaren te produceren, en ze leren de functies van verschillende soorten gebaren, zoals de deiktische (i.e., verwijzen naar een object of persoon) of verzoek (i.e., een object in handen krijgen) functies van aanwijs- of reikgebaren (Blake et al., 1994). Kinderen kunnen deze communicatieve gebaren vervolgens gebruiken om gezamenlijke aandacht te vestigen, specifieke informatie te verkrijgen en van volwassenen te leren (zie Tamis-LeMonda et al., 2014). Ten slotte zou de relatie tussen contingente interacties en het leren van taal kunnen worden beïnvloed door de toegenomen aandacht van het kind voor de reactie van de ouder (Chen et al., 2021; Kuhl., 2007; Masek et al., 2021a). Ondanks meer dan een halve eeuw onderzoek op dit gebied zijn er nog veel vragen onbeantwoord vanwege de grote complexiteit van het analyseren van ouder-kind interacties.

Ten eerste hebben bestaande onderzoeken zich grotendeels gericht op verbale reacties van ouders, terwijl communicatie **multimodaal** is. Het is vaak gebleken dat de temporeel en semantisch contingente reacties van ouders op de vocalisaties en gebaren van baby's de latere woordenschat van de kinderen kunnen voorspellen (e.g., Bornstein et al., 2008; McGillion et al., 2013; Olson & Masur, 2015; Tamis-LeMonda et al. ., 2001; Wu & Gros-Louis, 2015). De temporele en semantische contingentie

van de reactie van de ouder zou het voor het kind gemakkelijker kunnen maken om nieuwe woorden aan objecten of gebeurtenissen te kunnen koppelen. Communicatie is echter inherent multimodaal: auditieve en visuele informatie wordt vaak gelijktijdig aangeboden. Wanneer een baby naar een pop wijst en de ouder/verzorger reageert: “Wat een mooie pop!”, terwijl de ouder/verzorger deze ook nog oppakt en aan het kind laat zien, krijgt het kind twee gelijktijdige signalen voor de naam van het voorwerp: een auditief en een visueel signaal. De combinatie van auditieve en visuele signalen maakt het taalaanbod minder ambigu, waardoor het voor het kind gemakkelijker zou kunnen worden om het woord “pop” te koppelen aan de juiste referent in de omgeving (Baldwin et al., 1996; Gogate et al., 2000). Toch houden onderzoeken naar de relatie tussen responsiviteit van ouders en de woordenschat van kinderen zelden rekening met de non-verbale of multimodale (d.w.z. gecoördineerde verbale en non-verbale) reacties van ouders. Multimodale reacties van ouders verminderen de referentiële ambiguïteit in het taalaanbod waardoor deze reacties het mogelijk makkelijker maken voor kinderen om nieuwe woorden te leren.

Ten tweede hebben veel onderzoeken zich alleen gericht op het gedrag van het kind of de ouder tijdens sociale interacties en hoe dit individuele gedrag verband houdt met de woordenschatontwikkeling van kinderen, terwijl sociale interacties **bidirectioneel** zijn. In veel onderzoeken is bijvoorbeeld het verband tussen de aanwijsgebaren van baby's en hun woordenschat onderzocht, zonder rekening te houden met de reacties van ouders/verzorgers op deze aanwijsgebaren (zie Colonnesi et al., 2010) of het verband tussen het taalaanbod van ouders en de woordenschat van kinderen zonder rekening te houden met het gedrag van de kinderen tijdens interacties (zie Anderson et al., 2021). Leren vindt plaats tijdens sociale interacties die zowel door het kind als door de ouder worden gevormd (zie Renzi et al., 2017). De vraag blijft daarom hoe kinderen en ouders gezamenlijk bijdragen aan de woordleerervaring. In een recente studie hebben Chen et al. (2021) getoond dat wanneer ouders objecten benoemen, kinderen meer aandacht hebben voor de objecten wanneer de woorden nog onbekend waren voor het kind. Wanneer het woord nog nieuw was, hadden de ouders ook de neiging het object vaker aan te raken tijdens het benoemen. Bovendien resulteerde dit,

wanneer ouders het voorwerp aanraakten tijdens het benoemen ervan, ook weer in langduriger kijkgedrag (i.e., aandacht) van het kind voor het object (Chen et al., 2021). Dit laat zien hoe het individuele gedrag van zowel het kind als van de ouder het gedrag van de ander weer beïnvloedt tijdens sociale interacties. In een ander onderzoek hebben Ger et al. (2018) ontdekt dat bepaalde details in het wijsgedrag van baby's, zoals de vorm van de handen of het maken van geluiden tijdens het aanwijzen, de reacties van ouders op de aanwijsgebaren van baby's weer beïnvloedden. Uit het onderzoek bleek ook dat de semantisch contingente reacties van ouders op de aanwijsgebaren van baby's op de leeftijd van 10 maanden gerelateerd waren aan een toename in het wijsgedrag van baby's op de leeftijd van 12 maanden. Het analyseren van het gezamenlijke gedrag van ouders en baby's helpt ons beter te begrijpen hoe beide bijdragen aan de leeromgeving, wat ons weer beter helpt te begrijpen hoe kinderen succesvol nieuwe woorden leren.

Het overkoepelende doel van deze dissertatie is het voorspellen van de woordenschat van Nederlandse kinderen in een grootschalig, longitudinaal onderzoek. We hebben gekozen voor een dyadische aanpak bij het bestuderen van verbale, non-verbale, en multimodale gedragingen tijdens ouder-kind interacties. Hiervoor hadden we een nieuw coderingsschema nodig waarin zowel verbale als non-verbale reacties van ouders/verzorgers waren opgenomen. We hadden ook betrouwbare metingen nodig van de woordenschat van Nederlandse kinderen. Deze dissertatie bestaat uit vier empirische artikelen die methodologische en theoretische vraagstukken beantwoorden die nodig zijn voor het behalen van het overkoepelende doel.

Hoofdstuk 2

Om ouder-kind interacties te kunnen bestuderen hebben we allereerst annotaties en transcripties nodig van deze interacties. Het handmatig annoteren en transcriberen van geluidsopnames en/of videobeelden is een enorm tijdrovende klus. Daarom hebben we in Hoofdstuk 2 onderzocht of we geautomatiseerde hulpmiddelen kunnen gebruiken bij dit handmatige proces. Zo zijn er momenteel veel automatische spraakherkenningssystemen beschikbaar. Zulke systemen zijn getraind op grote

hoeveelheden volwassengerichte spraak (d.w.z. spraak door volwassenen gericht aan andere volwassenen). Dit zou weleens tot problemen kunnen leiden wanneer het doel is om kindgerichte spraak te herkennen. Kindgerichte spraak wordt akoestisch gekenmerkt door een hogere toon, een groter toonbereik en een langzamere spraaksnelheid vergeleken met volwassengerichte spraak. Deze akoestische verschillen zouden kunnen leiden tot meer herkenningsproblemen wanneer een systeem is getraind met volwassengerichte spraak, maar dit is nog niet eerder onderzocht.

In Hoofdstuk 2 hebben we de nauwkeurigheid van bestaande spraakherkenningsystemen voor het transcriberen van kindgerichte spraak onderzocht. In het eerste experiment hebben we vergeleken hoe nauwkeurig Kaldi-NL, een open-source spraakherkenningssysteem, specifieke naamwoorden kon herkennen in volwassengerichte spraak en kindgerichte spraak. We hebben gevonden dat maar 55,8% van de woorden in kindgerichte spraak correct werden herkend, terwijl 66,8% van de woorden in volwassengerichte spraak correct werden herkend. Er waren significante negatieve effecten van spraakregister (kindgerichte spraak) en toonhoogte op de nauwkeurigheid. Gezien de vrij lage prestaties van het bestaande spraakherkenningssysteem op beide spraakregisters, hebben we in het tweede experiment breder onderzocht hoe accuraat volledige zinnen worden herkend. Dit hebben we gemeten door alle kindgerichte spraak te transcriberen en foutpercentages van volledige zinnen, in plaats van alleen de specifieke naamwoorden, te berekenen. We hebben in dit tweede experiment bovendien twee open-source spraakherkenningssystemen met elkaar vergeleken. Bij het transcriberen van volledige zinnen had Kaldi-NL een foutpercentage van 40,1% terwijl het nieuwe systeem WhisperX maar een foutpercentage had van 22,5% tijdens het herkennen van volledige zinnen in kindgerichte spraak. Dit laatste systeem is accuraat genoeg om de handmatige annotatieprocedure van kindgerichte spraak te vergemakkelijken. Afhankelijk van de doelen van het onderzoek moeten deze automatische transcripties handmatig worden verbeterd, maar dit zal een stuk vlotter gaan dan wanneer de transcripties volledig handmatig worden gemaakt.

Hoofdstuk 3

Voordat we variatie in de woordenschat van Nederlandse kinderen kunnen onderzoeken, hebben we eerst woordenschatmetingen nodig die betrouwbaar en valide zijn en genoeg onderscheid tussen kinderen kunnen maken. In Hoofdstuk 3 hebben we de validiteit en betrouwbaarheid van de NYOUth-CDI's van meer dan 300 Nederlandse kinderen bekeken voordat we ze gingen gebruiken om individuele verschillen tussen kinderen te bestuderen. Vervolgens hebben we onderzocht of bekende demografische voorspellers van variatie, zoals het opleidingsniveau van de moeder of het geslacht van het kind, leeftijdsspecifiek en/of taakspecifiek zijn in deze grote, longitudinale steekproef. Slechts een beperkt aantal onderzoeken heeft de effecten van deze voorspellers longitudinaal onderzocht op verschillende woordenschatmetingen gedurende de ontwikkeling van het kind, terwijl er nog onbeantwoorde vragen zijn over het beginmoment en de stabiliteit van deze effecten door de tijd heen.

Hoofdstuk 3 laat zien dat de NYOUth-CDI's een hoge betrouwbaarheid en sterke validiteit hebben. Dit suggereert dat de data geschikt zijn om de variatie in de woordenschat van Nederlandse kinderen te kunnen bestuderen. De resultaten laten zien dat voor baby's rond de leeftijd van 10 maanden de gebarenschaal (die de omvang van het gebarenrepertoire van het kind aangeeft) hogere validiteit heeft vergeleken met de twee woordenschatschalen (woordproductie en woordbegrip). Het gebarenrepertoire van baby's, gerapporteerd door de ouders d.m.v. een checklist, was de enige maat op deze leeftijd die significant gecorreleerd was met de receptieve woordenschat van kinderen enkele jaren later. Vervolgens onderzochten we de effecten van de bekende demografische voorspellers van variatie in woordenschat. Hoewel we ontdekten dat het geslacht van de kinderen, het opleidingsniveau van de moeder en meertaligheid een deel van de variatie in de woordenschat van kinderen konden verklaren, waren de effecten leeftijdsspecifiek en taakspecifiek. Dit suggereert dat het handig is om de effecten van mogelijke voorspellers van woordenschat te meten op verschillende leeftijden, door middel van verschillende woordenschattaken. Dit zal de validiteit van onderzoeksresultaten vergroten.

Hoofdstuk 4

In Hoofdstuk 4 presenteerden we een karakterisering van de vocalisaties en gebaren van baby's rond de 10 maanden oud en de verbale, non-verbale en multimodale reacties van de ouders gedurende zes minuten vrij spel tussen ouder en kind. Tot nu toe zijn de non-verbale kanten van responsiviteit van ouders/verzorgers grotendeels onderbelicht gebleven in voorgaande studies. We hebben daarom een nieuw coderingsschema met verschillende categorieën van verbale reacties, gebaren, gezichtsuitdrukkingen en andere lichamelijke reacties van ouders ontwikkeld en getest. We onderzochten tevens of de verschillende gedragingen van baby's de neiging hadden om verschillende soorten reacties van ouders uit te lokken.

Ten eerste ontdekten we dat ouders een reeks verbale, non-verbale en multimodale gedragingen gebruiken wanneer ze reageren op de vocalisaties en gebaren van baby's rond de 10 maanden oud. Hoewel de overgrote meerderheid van de reacties verbaal waren (d.w.z. gesproken taal) was ca. 40% van deze verbale reacties multimodaal. Dit betekent dat ze op zijn minst gedeeltelijk overlappend waren met non-verbaal gedrag, bijvoorbeeld een handgebaar. Vervolgens ontdekten we dat verschillende gedragingen van baby's verschillende responspercentages en typen reacties van ouders neigen uit te lokken. De resultaten laten zien dat bimodaal gedrag van baby's (d.w.z. een combinatie van een vocalisatie en handgebaar) meer verbale en multimodale reacties van ouders uitlokte, terwijl de gebaren van baby's hogere percentages non-verbale reacties uitlokten. Ouders reageerden het minst vaak op vocalisaties die niet gepaard gingen met een handgebaar. We vonden ook dat wijsgebaren van kinderen meer verbale reacties van ouders uitlokten, terwijl speelgoed aangeven meer non-verbale en multimodale reacties van ouders uitlokten. Deze bevindingen suggereren dat baby's zelf een rol kunnen spelen bij het vormen van hun vroege leeromgeving door specifieke gebaren of geluiden te maken, waardoor ze de responsiviteit van hun ouders/verzorgers kunnen beïnvloeden.

Hoofdstuk 5

Ten slotte hebben we in Hoofdstuk 5 gekeken of dyadische gedragingen (gecombineerde gedragingen van baby's met de uitgelokte reacties van ouders/verzorgers) betere voorspellers waren van de woordenschat van kinderen dan het individuele gedrag van de baby's. Om dit te onderzoeken hebben we de relatieve voorspellende waarde van drie subsets van voorspellers op de woordenschat van kinderen vergeleken: 1) de frequentie van individuele gedragingen van baby's ongeacht de reactie van ouders/verzorgers, 2) de frequentie van gedragingen van baby's die een verbale reactie van ouders/verzorgers hebben uitgelokt en 3) de frequentie van gedragingen van baby's die een multimodale reactie van ouders/verzorgers hebben uitgelokt. We onderzochten het gedrag van baby's rond de 9–11 maanden en gebruikten deze om de gelijktijdige en latere woordenschatuitkomsten, gemeten rond de leeftijd van 2–4 jaar, van de kinderen te voorspellen. Geen enkele eerdere studie heeft verschillende combinaties van dyadische gedragingen, met verschillende typen reacties van ouders, direct met elkaar vergeleken. Het was nog onontdekt of we verschillende effecten vinden bij het koppelen van verschillend babygedrag (vocalisaties, wijzen, laten zien + aangeven) aan verschillende soorten reacties (verbaal of multimodaal) van ouders.

We vonden dat de frequentie van de wijsgebaren van baby's verband hield met de latere receptieve woordenschat van de kinderen, terwijl de frequentie van laten zien en aangeef gebaren van baby's (d.w.z. de baby heeft speelgoed vast en laat het zien of geeft het aan de ouder) verband hield met de latere productieve woordenschat van de kinderen – maar alleen wanneer we de specifieke instanties die multimodale reacties van ouders/verzorgers hadden uitgelokt tijdens de sociale interactie meenamen in de analyse. Dit komt mogelijk doordat kinderen makkelijker kunnen leren van multimodale reacties We kunnen stellen dat de dyadische combinaties van gebaren van baby's en verbale en multimodale reacties van ouders een grotere voorspellende waarde hebben voor de woordenschat van kinderen dan het individuele gedrag van baby's. Dit benadrukt het belang van het bestuderen van dyadische gedragingen tijdens sociale interacties tussen ouder en kind. Tevens hebben we voor

het eerst laten zien dat multimodale reacties van ouders uniek bijdragen aan het vergroten van de voorspelbaarheid van kind gedragingen op hun woordenschatuitkomsten.

Conclusie

Het doel van dit proefschrift was om de variatie in de woordenschat van Nederlandse kinderen te voorspellen door ouder-kind interacties te onderzoeken in een grote, longitudinale cohortstudie. Hoewel er momenteel grote belangstelling is voor het annoteren van kindgerichte spraak, zoals ook nodig was in dit proefschrift, was de nauwkeurigheid van bestaande automatische spraakherkenningssystemen voor de annotatie van kindgerichte spraak nog onbekend. We hebben aangetoond dat onderzoekers met succes bestaande automatische spraakherkenningssystemen kunnen gebruiken om het arbeidsintensieve handmatige annotatieproces te vergemakkelijken. Ten tweede hebben we, na het vaststellen van de validiteit en betrouwbaarheid van de NYOUth-CDI's, aangetoond dat bekende demografische voorspellers van de woordenschat van kinderen leeftijdsspecifiek en taakspecifiek zijn. Dit benadrukt het belang van het opnemen van meerdere uitkomstmaten voor het bestuderen van voorspellers van variatie in de woordenschat van kinderen. Ten derde benadrukken de resultaten het belang van het bestuderen van dyadische en multimodale gedragingen bij het onderzoeken van ouder-kind interacties. Tijdens interacties vormen baby's en ouders elkaars gedrag voortdurend. Wanneer we alleen het gedrag van de ouder of het gedrag van het kind los ven elkaar bestuderen, mist de helft van het plaatje. Daarbij is het ook belangrijk om multimodale aspecten van interacties te bestuderen. Visuele signalen in combinatie met gesproken taal kunnen kinderen helpen om referentiële ambiguïteiten in het taalaanbod op te lossen. In onderzoeken naar de ontwikkeling van de woordenschat van kinderen willen we beschrijven hoe baby's voldoende informatie uit het taalaanbod halen die hen in staat stelt om woorden te leren. Het bestuderen van de dyadische en multimodale aard van sociale interacties tussen ouder en kind schetst een completer beeld van de vroege leeromgeving van kinderen, wat ons weer een stapje dichter bij het oplossen van deze puzzel brengt.

Curriculum Vitae

Anika van der Klis was born on October 12th, 1994 in Woerden, the Netherlands. From 2014 to 2017, she studied English Language and Culture at Leiden University where she specialised in linguistics and second language acquisition. She also completed the extracurricular Honours programme. In her third year, she participated in the Research Traineeship of the Faculty of Humanities where she completed a research project under the supervision of prof Claartje Levelt and prof Lisa Cheng. She wrote her Bachelor's thesis under the supervision of dr Bert Botma (grade 9.0). In 2017, she started the Research Master's degree in Linguistics at Leiden University. She completed courses related to language acquisition, cognitive neuroscience, bilingualism, phonetics, and phonology. During her Master's, she also worked as a research assistant for the University of Amsterdam. Her Master's thesis was supervised by prof Claartje Levelt and dr Josje Verhagen (grade 8.0). In 2019, she started her PhD project at the Faculty of Humanities of Utrecht University under the supervision of prof René Kager and dr Frans Adriaans. Since November 2023, she is working as a postdoc at the Faculty of Social and Behavioural Sciences of Utrecht University with dr Caroline Junge.

List of publications

van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2020). Automatic recognition of target words in infant-directed speech. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 522. https://doi.org/10.1145/3395035.3425184

van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2023). Using open-source automatic speech recognition tools for the annotation of Dutch infant-directed speech. *Multimodal Technologies and Interaction*, *7*(7). https://doi.org/10.3390/mti7070068

van der Klis, A., Adriaans, F., & Kager, R. (2023). Infants' behaviours elicit different verbal, nonverbal, and multimodal responses from caregivers during early play. *Infant Behavior and Development*, *71*, 101828.https://doi.org/10.1016/j.infbeh.2023.101828

van der Klis, A., Kaya, H., Najafian, M., & Safavi, S. (2022). 3rd ICMI Workshop on Bridging Social Sciences and AI for Understanding Child Behaviour. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 807–809. https://doi.org/10.1145/3536221.3564031

van der Klis, A., van Lieburg, R., Cheng, L. L., & Levelt, C. C. (2023). Pauses matter: Rule-learning in children. *Language Development Research*. https://doi.org/10.34842/2023.0466