The production-comprehension link in prosodic development and individual differences

Aoju Chen and Huub van den Bergh

Utrecht University

Acknowledgements

This study is supported by a VIDI grant (276-89-001) awarded to the first author by the NWO (Netherlands Organisation for Scientific Research). We thank the children and teaching staff from Houten Montessori Primary School, De Ontdekkingsreis Primary School (Doorn), Soest Montessori Primary School, De Wegwijzer Primary School (Soest) for their indispensable cooperation in this research.

Abstract

Central to the debate on the production-comprehension link in prosodic development is the acquisition of focus-to-prosody mapping. To elucidate the nature of the production-comprehension link and shed first light on individual differences in the prosodic domain, the present study investigated developmental changes in production and comprehension of the focus-to-prosody mapping in Dutch-speaking children (age range: 4;8 ~ 7;5, N = 71) longitudinally. It was found that children's comprehension is predictive of their production only if their comprehension is already adult-like but their production isn't. Notably, individual differences in the production-comprehension link change with both sentence-position and age, challenging the assertion in the literature that individual differences are stable across development and domains in first language acquisition.

1. Introduction

In theories of language acquisition, comprehension is considered the prerequisite for production (e.g., Clark 1993, Hendriks and Koster 2010). Comprehension should therefore precede production or at least develop in tandem with production. Language development indeed generally follows these tendencies. However, in certain areas of language development production seems to precede comprehension, known as the production-precedes-comprehension asymmetries (see Hendriks and Koster 2010 for review). A case in point is the acquisition of sentence-level speech prosody (hereafter prosody, referring to melody and rhythm in speech) in West Germanic languages. Past work has shown that children have largely acquired the adult inventory of prosodic contours and can use some of the prosodic properties communicatively at the age of 2 or 3, but they do not yet process prosodic information in language comprehension as efficiently as adults do at the age of 4 or 5, and fail to interpret certain uses of prosody even at the age of 9 or 10 (e.g. Cruttenden 1985; Cutler and Swinney 1987).

Central to the discussion on the production-comprehension link in prosodic development is the interface between prosody and focus. Focus refers to the predication on a topic in a sentence and typically contains new or contrastive information to the hearer (Lambrecht 1994; Vallduví and Engdahl 1996). In many languages, there is a strong association between the focal constituent and prosodic prominence and between post-focal constituents and a reduced or little prosodic prominence (e.g., Bolinger 1978; Xu 1999; Xu and Xu 2005). In West Germanic languages, prosodic prominence can be varied through placement of pitch accent (e.g. accenting a word or not), choice of accent type (e.g. a falling pitch contour vs. a rising pitch contour) and phonetic realisation of a pitch accent (hereafter phonetic realisation) (e.g. realising a falling pitch accent with a larger or smaller pitch range) (Gussenhoven 2004, Chen 2012). Speakers tend to accent the focused constituent, typically

with a falling accent, and deaccent the post-focus constituents or accent them with a downstepped falling accent (i.e. a falling accent with a lowered pitch peak relative to the pitch peak of the preceding accent) (Ladd 2006). Listeners take the focus-to-prosody mapping into account in online language comprehension such that an appropriate focus-to-prosody mapping speeds up comprehension, compared to an inappropriate focus-to-prosody mapping (Birch and Clifton 1995, Cutler et al. 1997). Relatedly, listeners use the broad mapping between information value (new vs. given) and prosody to predict whether the upcoming referent is a new referent or a previously mentioned referent in online reference resolution before the segmental information completely unfolds (e.g. Dahan et al. 2002, Weber et al. 2006, Chen et al. 2007, Ito and Speer 2008, Watson et al. 2009). Adult-like competence in prosodic focus-marking thus entails that children can both vary prosodic prominence to encode focus in production and efficiently exploit the focus-to-prosody mapping in language comprehension.

It is widely believed that children under the age of 6 can use prosodic prominence to mark focus in production but fail to interpret or efficiently use the focus-to-prosody mapping in comprehension (Cruttenden 1985; Szendrői, 2004; Hendriks 2005; Müller et al. 2006). This production-precedes-comprehension asymmetry would mean that children can capitalise on the relevant linguistic knowledge of prosodic focus marking in production but not in comprehension. Cutler and Swinney (1987), however, postulated that 4 to 5-year-olds' success in production was not driven by linguistic knowledge of prosodic focus marking, similar to their comprehension. They argued that 4- to 5-year-olds' production was merely a universal 'physiological reflex', whereby speakers increase pitch and thus prosodic prominence as a result of excitation when telling something new or important (Bolinger 1983). They thus concluded that there was no evidence for a production-precedes-comprehension asymmetry in the acquisition of prosodic focus-marking in 4- to 5-year-olds.

Recent cross-linguistic studies of children's prosodic focus-marking in production have shown that children use prosody in different ways in different languages for the purpose of focus marking already at the age of 4 or 5 (see Chen 2018a for a review). Importantly, their use of prosody is subject to factors such as the relative importance of various prosodic cues to focus, the acoustic salience of the cues in perception, and transparency of the form-function between a prosodic cue and focus in the target language. These findings suggest that 4- to 5-year-olds' use of prosody cannot be explained by Bolinger's universal, which would predict similarities in their prosodic focus marking across languages. Instead, 4- to 5-year-olds are likely to have acquired some linguistic knowledge of prosodic focus marking in the target language.

Do 4- to 5-year-olds really perform poorly then in tasks on comprehension of the focus-to-prosody mapping? Chen (2010) reviewed the comprehension studies cited as evidence for children's failure in comprehension in the literature and found that these studies did not directly examine children's comprehension of the focus-to-prosody mapping but were concerned with children's ability in uses of prosodic prominence other than highlighting the focal word. Consequently, children's poor performance on the comprehension tasks in the earlier comprehension studies does not necessarily imply their inability to process the focus-to-prosody mapping. Besides, the test materials used in the earlier comprehension studies were usually syntactically more complex and semantically more demanding (e.g. SV + direct Object + indirect Object sentences with the particle 'only' or compound sentences containing ambiguous pronouns) than the materials used in the production studies (e.g. SVO or SV sentences). The apparent differences in syntactic complexity between the materials used in comprehension studies and those used in production studies have been suggested as a possible account for the production-precedes-comprehension asymmetries reported for other areas of language development (Hendriks and Koster 2010).

Chen (2010) subsequently examined 4- to 5-year-olds' and adults' semi-spontaneous production and online comprehension of the focus-to-prosody mapping in Dutch SVO sentences. The production data were obtained via an interactive picture-matching task. In this task, children answered the experimenter's questions on some pictures to help the experiment to pair them with other pictures. The comprehension task was a mirror image of the picture-matching task. In this task, a boy asked his pets questions on some pictures and children judged whether the pets gave good answers to the questions. The answers were spoken with prosody either appropriate or inappropriate in terms of focus marking. It was found that children were similar to adults in both production, as revealed by analysis on accent placement in focus and non-focus, and comprehension, as revealed by children's reaction times in the two prosodic conditions. The differences between children and adults were of a gradient nature, i.e. more frequent use of accentuation in post-focus constituents and slower reaction times in general in children than in adults. In the light of these results, Chen (2010) argued that there was no evidence for an asymmetry in the acquisition of the focus-to-prosody mapping in 4- to 5-year-old Dutch-speaking children when syntactic complexity and task complexity were comparable in the production and comprehension experiments.

However, the debate on the production-comprehension link in the acquisition of prosodic focus-marking is far from being settled yet for several reasons. Firstly, if comprehension is the prerequisite for production, children's comprehension is expected to become adult-like at an earlier age than their production or at least at the same age when their production becomes adult-like, and development in comprehension should be predictive of development in production. Past work is exclusively concerned with the temporal order in which children obtain adult-like production and comprehension, similar to this line of research in other aspects of language acquisition, e.g. acquisition of pronouns and reflexives and acquisition of SVO word order (Hendriks and Koster 2010 for review). The question of

whether development in comprehension can predict development in production has not been addressed. It is thus of theoretical importance to gain insight into the relation between production and comprehension in addition to the temporal order in which they reach the adult-like level.

Secondly, production studies of prosodic focus marking in children acquiring a West-Germanic language have focused on the use of accent placement. Although accent placement is an important device for focus marking, accent type and phonetic realisation also matter. For example, a focus constituent can be accented with both a fall and a downstepped fall in sentence-final position but the fall occurs more frequently than the downstepped fall in adult Dutch (Chen 2011). When the same accent type is used in both focus and non-focus, as found in sentence-initial subject noun phrases, adult speakers realise the falling accent in the focus condition with a larger pitch range and a longer word duration than in the non-focus condition (Chen 2009). Dutch-speaking children become adult-like in choice of accent type in SVO sentences in sentence-initial position at the age of 4 or 5 and in sentence-final position at the age of 7 or 8 (Chen 2011). They do not use phonetic realisation to distinguish focus and non-focus in sentence-initial subject nouns and only vary pitch range of the falling accent at the age of 7 or 8 (Chen 2009). Non-adultlike use of phonetic implementation has perceptual consequences. In a perception study, adult native speakers of Dutch performed at chance level when asked to determine whether a sentence-initial NP (e.g., *de paashaas* ‘the Easter bunny’) was the subject of a sentence answering a who-question (e.g. Who paints the football?) or a what-question (e.g. What does the Easter bunny paint?) in 4- to 5-year-olds’ speech. Although their performance improved when presented with speech from 7- to 8-year-olds (a 64% accuracy), the highest accuracy was obtained when presented with speech from adults (a 76% accuracy) (Chen 2018b). These findings call for the need to go beyond

analysing the distribution of accents in research on the production-comprehension link in the acquisition of prosodic focus-marking.

Thirdly, past work was primarily concerned with children at the age of 4 to 5 years. Recent studies of children's prosodic focus-marking in production have shown notable development in various aspects of prosodic focus marking after the age of 5 (Chen 2018a). This raises the question of whether the relation between production and comprehension undergoes changes over time.

Lastly, existing literature on the production-comprehension link in prosodic development is based on studies examining production in one group of children and examining comprehension in another group of children. Little is known on whether the production and comprehension are related in the same way for different children. Research on individual differences in first language acquisition is mostly concerned with vocabulary (Frank et al. 2017) and grammar development (Bishop 2003). It has been argued that individual differences are 'large and notably stable across development ... and across all domains' in first language acquisition (Kidd et al. 2018). Wells, Peppé and Goulondris (2004) found that individual differences could occur at both younger and older ages and in different aspects of prosodic development in their study with 120 English-speaking 7- to 13-year-olds performing multiple production and comprehension tasks, consistent with Kidd et al.'s (2018) claim. In contrast, production data from a sample of 22 4- to 5-year-old Dutch-speaking children and 18 7- to 8-year-olds showed that there were noticeable individual differences in prosodic focus marking in the younger children but not in the older children (Chen 2011). It thus remains to be investigated what the range of individual differences is like in production and comprehension of prosodic focus marking and in the relation between production and comprehension.

In the current study, in order to obtain a comprehensive picture on the production-comprehension link in acquisition of prosodic focus marking and individual differences in this respect, we examined prosodic focus marking in over 70 Dutch-speaking children of different ages longitudinally and addressed three questions: (1) what is the development like in production and comprehension between 4 and 8 years of age? (2) Is children's comprehension predictive of their production? (3) How do children differ in production, comprehension and the relation between production and comprehension? We were concerned with narrow focus (i.e. focus on one lexical word) in sentence-initial position and narrow focus in sentence-final position in SVO sentences, because prosody is used in different ways to encode focus in these two sentence-positions in Dutch. Children were tested several times at regular intervals in a period of eight months. Production data were obtained from children and an adult control group in a natural and interactive setting via an adjusted version of the method used in Chen (2011). The sentences produced by children and adults were subsequently evaluated in a perceptual rating experiment, in which trained adult listeners rated the appropriateness of the prosody of each sentence in the corresponding context on a five-point scale. The scores for prosodic appropriateness reflected the overall use of prosody, not just accent placement, and factored the perceptual relevance of details in production in. Comprehension was examined using the same method as in Chen (2010). We tested a set of hypotheses derived from the above-reviewed work on Dutch-speaking children's prosodic focus marking reviewed above. The hypotheses concerned development in children's production and comprehension (question (1)), the production-comprehension link (question (2)), and individual differences (question (3)).

Regarding production, because children are adult-like in the use of accent placement and accent type but are not in the use of phonetic realisation by the age of 8 and the use of phonetic realisation has perceptual consequences (Chen 2009, 2011, 2018b), we have

hypothesised that sentence-position matters to the age at which children's production is perceived to be target-like (**Hypothesis 1a**). Our prediction was that children's production is perceived to be target-like at an earlier age in final focus than in initial focus. Regarding comprehension, in line with Chen (2010), we hypothesised that children are adult-like in comprehension at all ages within the age-range tested (**Hypothesis 1b**). Our prediction was that regardless of age, children respond faster on the average when the focus-to-prosody mapping is appropriate than when it is inappropriate. Regarding the relation between production and comprehension, consistent with our predictions for production and comprehension, we hypothesised that children's comprehension outperforms production throughout the age range tested here in sentence-initial focus and only at younger ages in sentence-final focus, in line with the comprehension-precedes-production asymmetry widely observed in language acquisition (**Hypothesis 2**). Our prediction was that children's comprehension is predictive of their production regardless of age in sentence-initial focus but the predictive power of comprehension for production interacts with age in sentence-final focus.

Lastly, in the light of findings on vocabulary and grammar development and findings from Wells et al. (2004), we hypothesised that individual differences exist to a similar degree in different focus conditions at all ages within the age-range tested regarding production, comprehension and the strength of the predicted relation between production and comprehension (**Hypothesis 3**). Bearing statistical modelling to be done in mind, individual differences were operationalised as differences in a prosodic ability between children at the mean age of all participating children (addressing the existence of individual differences) and differences in changes in a prosodic ability with age and with different sentence-positions between children (addressing the stability of individual differences). Our prediction was thus

that children differ in prosodic abilities at the mean age but not in the degree of changes in their prosodic abilities with age and in different sentence positions.

2. The method

2.1 Participants

Seventy-one Dutch-speaking children from four age groups participated in the longitudinal study as part of a large-scale research project on prosodic development. The children were on average 4;8, 5;6, 6;6 and 7;5 in their respective age group at the start of the longitudinal study. Detailed information on age and gender can be found in Table 1, in which the four groups of children are labelled as a4, a5, a6, and a7 according to their age at the start of the longitudinal study. The children were from monolingual Dutch-speaking families and were recruited from four primary schools in the Province of Utrecht in the Netherlands. They had no hearing loss and speaking disability according to parents' reports. We received written informed consent from the parents for the children to participate in the longitudinal study and for their experimental sessions to be recorded and filmed.

The children were tested at three longitudinal moments at intervals of about four months. Only at the first and third longitudinal moments, they did both the production and comprehension experiment. Data from these two longitudinal moments were analysed in the current study. For clarity's sake, these two longitudinal moments are referred to as the first testing moment and the second testing moment in this report. Given the comparability of the speech materials in the production and comprehension experiments, we conducted the production experiment first and then the comprehension experiment at each testing moment avoid potential influence of how prosody was used in the comprehension experiment on children's use of prosody in the production experiment.

Table 1. Mean ages, age ranges and gender of the four groups of children at the two testing moments (time 1, time 2) in the longitudinal study.

Group_time1	total	Production		comprehension	
		age_time1	age_time2	age_time1	age_time2
a4	12 (5 m)	4;8 (4;2-4;12)	5;3 (4;10 - 5;7)	4;8 (4;2 - 5;1)	5;4 (4;10 - 5;8)
a5	26 (12 m)	5;6 (5;0 - 5;12)	6;2 (5;8 - 6;8)	5;7 (5;0 - 6;0)	6;3 (5;8 - 6;8)
a6	15 (7 m)	6;6 (6;1 - 6;12)	7;2 (6;9 - 7;7)	6;7 (6;1 - 6;12)	7;2 (6;9 - 7;8)
a7	18 (10 m)	7;5 (6;12 - 7;12)	8;1 (7;7 - 8;8)	7;6 (7;0 - 8;1)	8;1 (7;8 - 8;8)

Note: Some children happened to be tested shortly before their birthday at a testing moment.

Their age was registered as X;12, instead of X+1 years.

An additional fifty-six children participated in the longitudinal study but their data were not included in the current analysis. The data of eight of these children were lost due to loss of equipment. The other children either did not participate in both experiments at both testing moments, or did not have production ratings because their production in the focus conditions of interest was not evaluated for reasons like poor articulation, stuttering, self-repair, and being accompanied by unexpected noise in their surroundings.

Twenty-three adult native speakers of Dutch (mean age 21;5, range: 18;8 – 28;8, 10 men, 13 women) participated in the production study as the control group. They were students at Utrecht University at the time of testing and were tested on the same tasks following the same procedure as the children. Prior to the experiment, they were informed that the tasks were of a simple nature as they were also used on child participants.

2.2 The production study

The production study consisted of a production experiment and a rating experiment. In the production experiment, SVO sentences were elicited in different focus conditions. Focus

conditions were varied by means of the question-answer paradigm (Roberts 1996). In the rating experiment, adult raters rated the appropriateness of the prosody in the participants' answer sentences in the corresponding contexts.

2.2.1 The production experiment

We adapted the picture-matching game used in Chen (2011) to elicit SVO sentences from the participants in a natural and interactive setting. The game was played in experimenter-participant dyads. Game materials included sets of hand-drawn colour pictures. The participant's task was to help the experimenter to find the matching picture for each of her pictures by answering her questions. The experimenter was thus dependent on the participant and the participant was in control of the game, making the game very appealing for children. The conversation between the experimenter and the participant was primarily composed of short question-answer dialogues. In Chen's (2011) game, the participants needed to seek information first from a virtual robot before answering the experimenter's questions. In the current version, the participants had all the information visually presented in their own set of pictures (more details on section 2.2.1.2).

Fifteen question-answer dialogues were embedded in the game to elicit 15 SVO sentences in five focus conditions: narrow focus in sentence-initial position (initial focus), responding to who-questions, and narrow focus in sentence-final position (final focus), responding to what-questions, narrow focus in sentence-medial position, responding to what-does-X-do-with-Y questions, contrastive focus in sentence-medial position, correcting the experimenter's statement about the action, and broad focus over the whole sentence, responding to what-happens questions. Only the first two focus conditions were relevant to the current study; the other three were included for other purposes. The target SVO sentences were unique combinations of five subject-nouns (*baker*, 'baker', *hond*, 'dog', *leeuw*, 'lion',

meisje, ‘girl’, *poes*, ‘cat’), three verbs (*tekenen*, ‘draw’, *koken* ‘cook’, *toveren* ‘conjure’), and three object-nouns (*lepel*, ‘spoon’, *laars* ‘boot’, *wortel* ‘carrot’). All words were highly familiar to Dutch 4-year-olds. The nouns in the initial focus condition were not the same nouns as the nouns in the final focus condition. This is, however, not an issue in the current study because the production was examined via adults’ rating, instead of acoustic analysis, where the occurrence of same words in different focus conditions would be necessary.

To make sure that the participants would use the intended words in the picture-matching game, we asked each participant to complete a picture-naming task prior to the game during the same session. In the picture-naming task, the participants were familiarised with the characters and actions that appeared in the picture-matching game (more in section 2.2.1.1). The characters were thus referentially given to the participants at the start of the picture-matching game, rendering the use of a definite article in reference in the picture-matching game appropriate.

Three female student assistants, who were native speakers of Dutch and students of linguistics or literature, conducted the production experiment with the children individually at their schools during school time and with the adults in a sound-attenuated booth in the Linguistics Laboratory at the Utrecht Institute for Linguistics. To ensure consistency of the highest degree between the experimenters, we trained the experimenters on both the procedure and the use of prosody using an experiment protocol before they started testing participants.

To help the children to feel at ease with the experimenters, every experimenter picked up the child that she was to test from the classroom and had a chat with the child prior to the experiment. The testing schedule was made in such a way that every child did both the production and comprehension experiment with the same experimenter at one or both testing moments.

2.2.1.1 The picture-naming task

In the picture-naming task, the participants first named the nouns illustrated in individual pictures. The experimenter showed one picture a time to the participant, and invited the participant to name it by saying *Dit is een ...* 'This is a ...'. In the case of incorrect naming (e.g., calling a baker a cook), the experimenter first acknowledged that the entity might look like what the participant had in mind, and then drew the participant's attention to distinctive features in the picture (e.g. the dough on the table) and suggested to the participant the intended label for the entity. In the second, the experimenter showed the participant pictures of the actions occurring in the game, and explained to him which action each picture depicted. Finally, to check whether the participant has got all the labels right, the experimenter went through all the pictures and asked the participant to name them once more. If the participant did not have difficulty with naming the nouns, he only named the actions for the second time.

2.2.1.2 The picture matching game

Three sets of pictures were used in the game, one set for the experimenter, one set for the participant, and one set consisting of matching pictures for the experimenter's pictures. The experimenter's and participant's pictures were piled up neatly in an order corresponding to the trial numbers. The third set of pictures were left unordered in a box. The experimenters' pictures lacked some or all the information about an event. The participant's pictures depicted complete events. The pictures in the box depicted only the information missing in the corresponding pictures of the experimenter. The experimenter explained to the participant how the game worked, and introduced two rules of the game: Never reveal his pictures to the experimenter; always respond with full sentences. The game consisted of fifteen trials, corresponding to the fifteen question-answer dialogues. Each trial was carried out in a fixed

number of steps (for illustration see Appendix 1). First, the experimenter took a picture from her set of pictures (e.g. a picture of a girl drawing something on a piece of paper), drew the participant's attention to the picture, and briefly described it by saying *Kijk! Het meisje. Het lijkt alsof het meisje iets tekent*. 'Look! The girl. It seems that the girl is drawing something.' She then asked the participant a question about the picture (e.g. *Wat tekent het meisje?* 'What is the girl drawing?') or made a guess about the missing information in the case of contrastive focus. The participant then took a picture from his own set of pictures, and identified the information requested by the experimenter. The experimenter repeated her question when possible. The participant then answered the question in an SVO sentence (e.g. *Het meisje tekent de lepel*. 'The girl is drawing the spoon.'). The experimenter thanked the participant for his information, looked for the picture of the spoon in the box, and handed over both pictures to the participant for his approval.

The game proper was preceded by five practice trials. If the participant gave elided answers or full-sentence answers containing non-target words during the practice trials, the experimenter reminded the participant of the rules of the game or the intended words. This intervention turned out to be necessary only in the case of a small number of the 4- to 5-year-olds. Most children provided full-sentence answers using the intended words right from the start. Each session lasted about 15-25 minutes, and was recorded at a sampling rate of 44.1 kHz with 16 bits resolution and filmed. The film recordings were used to check consistency in how the game was carried out.

2.2.2 Data annotation

The audio recording of each participant was first orthographically annotated using Praat (Boersma 2001). Second, full-sentence responses were selected as usable responses if they were not plagued by any of the following factors: self-correction, use of pronouns, use of

non-target words, detectable hesitation-induced silences, responding to a non-target question, elided responses, overlap with the experimenter's speech, and poor recording quality. Third, the usable full-sentence responses and the corresponding questions or statements were selected and extracted as individual .wav files.

2.2.3 The rating experiment

The usable full-sentence responses and corresponding questions or statements were combined into context-response dialogues with a 300-ms interval between the question or statement and the response in each dialogue and a 1000-ms interval between dialogues. The three native speakers of Dutch who administered the production experiment served as the raters and rated each response in each dialogue on how well its prosody fitted in the context on a five-point Equal Appearing Interval scale, with 1 standing for 'does not fit' for and 5 standing for 'fits perfectly'. Prior to the rating, the raters were given sound examples illustrating what the prosody typically sounded like in each focus condition and written instructions on how to do the rating. They did the rating at least ten months after the production experiment at the second testing moment was completed. These raters were familiar with child speech and the speech material to be evaluated and were thus expected to be able to focus on the prosody in their rating instead of being distracted by other features such as the speaking rate, voice quality and articulation.

To minimise variation in the scores due to comparisons between children, the dialogues were presented to the raters per child and per age group. The rating experiment was conducted in 7 30- to 40-minute sessions. The raters could listen to each dialogue maximally three times before finalising the score. As only initial focus and final focus were relevant to us here, the scores for these two focus conditions were included for further analysis.

2.3 The comprehension study

2.3.1 The correct-incorrect answer game

To examine children's processing of the prosody-to-focus-mapping, we conducted the reaction-time (RT) experiment used in Chen (2010) on our participants individually about seven to ten days after the production experiment at each testing moment. Chen's (2010) experiment was modelled on Birch and Clifton's RT paradigm. Birch and Clifton (1995) found that when adult native speakers of English were asked to judge whether the answer made sense in a question-answer dialogue (e.g. Isn't Kerry pretty smart? Yes, she teaches math.), they had shorter reaction times in answer sentences with appropriate prosody (e.g. Yes, she teaches MATH) than in answer sentences with inappropriate prosody (e.g. Yes, she TEACHES math). The main difference between the appropriate and inappropriate prosody lay in which word was accented. The RT technique has been widely used in studies of sentence processing in 4-year-olds and older children in recent years (Clahsen 2008). The 'make sense' task in Birch and Clifton's study was, however, believed to be too abstract for 4- to 5-year-olds, and was replaced with a 'correct-incorrect answer' task, in which listeners were asked to judge whether the answer was correct or incorrect in each dialogue. Similar to the picture matching game, the participants were in control of the game.

The experiment was presented to our participants as a 'correct-incorrect answer' game. In the game a boy looked through some pictures with his three pets, a parrot, a chicken, and a duck. The boy wanted to know whether his pets knew the pictures well and which of the pets knew the pictures best. To find this out, the boy showed one picture a time to one of his pets and asked the pet a question about the picture (see Appendix 2 for illustration). The participant could follow the question-answer dialogues between the boy and the pets via a headphone set and see the boy, his pets and the pictures via a computer screen. The participant's task was to judge whether the pets' answers were correct or incorrect

(‘goed’ or ‘fout’ in Dutch) by pressing the green button for ‘correct’ and the red button for ‘incorrect’ on the responding device (i.e. the pushbutton box or the keyboard. The keyboard was used 6 months after the start of the longitudinal study when the push-button box was no longer supported on experiment laptops at the Linguistics labs of the Utrecht Institute of Linguistics. No differences in reaction time were observed between the children who did the RT experiment using the push-button box and the children using the keyboard.

Two focus conditions were embedded via wh-questions in the game, i.e. initial focus and final focus, as was done in the picture-matching game. The two focus conditions were combined with two prosody conditions (appropriate focus-to-prosody mapping vs. inappropriate focus-to-prosody mapping), forming four experimental conditions: initial focus-appropriate prosody, initial focus-inappropriate prosody, final focus-appropriate prosody, and final focus-inappropriate prosody, as illustrated in Appendix 2. Following Birch and Clifton (1995), Chen (2010) defined inappropriate prosody in one focus condition as the appropriate prosody in the other focus condition. Subsequent analysis showed that the sentences with similar prosodic patterns were indeed comparable in pitch in the subject and object nouns and in mean sentence duration.

Each experimental condition was implemented on 24 question-answer dialogues, comparable to the dialogues occurring in the picture-matching game. The answers in the experimental dialogues were correct answers in terms of the segmental and lexical content. In addition, twenty question-answer dialogues were included as fillers. The answers in the fillers contained either a lexical error or a pronunciation error. Each experimental condition was implemented on twelve fillers with a lexical error, e.g. *eend* ‘duck’ instead of *kip* ‘chicken’, and eight fillers with a pronunciation error, e.g. *jaangen* instead of *jongen* ‘boy’. The experimental dialogues and fillers were derived from recordings of a male native speaker of Dutch (, who produced the questions in child-directed speech (Burnham, Kitamura, Vollmer-

Conna 2002), and a female native speaker of Dutch who produced the answers in her usual manner.

A Latin square was used to distribute the experimental dialogues and fillers over the four conditions, resulting in four lists, in each of which each answer sentence only occurred once but each condition was realised six times in the experimental dialogues, three times in the fillers with lexically wrong answers and twice in the fillers with wrongly pronounced answers. In total, every participant was presented with forty-four dialogues.

2.3.2 Procedure

The children did the experiment individually in a quiet room at their schools during school time. The experimenters who conducted the production experiment also conducted this experiment by means of the Zep Experimental Control Application (Veenker 2013). The exact list and stimulus order that a child got were randomly chosen by Zep. An approximately equal number of children were assigned to each stimulus order of each list. Each session lasted about 20-25 minutes starting with a practice session. In the practice session, the children were familiarised with the task and trained to press buttons or keys in a proper manner. Each session was filmed; the recordings were used to check consistency in procedural details. The same measures were taken to help the children feel at ease with the experimenter as in the production experiment (section 2.2.1).

The timeline of a trial was as follows: A target picture appeared together with the picture of the boy and one of his pets on the screen. At the same time, the boy said *Kijk* ‘look’ as an attention getter. Eight-hundred milliseconds later, he named an entity in the picture (e.g. *Een varken* ‘A pig’). The eight-hundred-millisecond delay was built in to allow the participants to take a proper look at the picture. Twelve-hundred milliseconds after the naming, the boy asked the question (e.g. *Wat wast de varken?* ‘What is the pig washing?’).

Twenty-two hundred milliseconds after the end of the question, the pet answered the question (*Het varken wast een bloes*. ‘The pig is washing a blouse.’). At the end of the answer sentence, a timer with one millisecond accuracy was activated and a picture of the buttons for correct and incorrect answers appeared on the screen to remind the children of pressing the buttons/keys.

The reaction time was recorded from the end of the answer sentence until a button/key was pressed; the correct–incorrect judgment was also automatically recorded. The children were instructed to press the button/key as quickly as they could, but not before the end of the answer sentence. A timeout device was set at four seconds after the end of a sentence. Two measurements were taken from each child, i.e. the ‘correct–incorrect’ judgement and the reaction time.

3. Statistical analysis and results

3.1 Statistical modelling

As our measurements (i.e. production scores and reaction times) from each testing moment were nested both within children and within items, we used multilevel modelling to analyse the production data, the comprehension data and the relation between production and comprehension (Quené and van den Bergh 2008; Goldstein 2011; Snijders and Bosker 1999). Our model can be written as the equation in (5) with $Y_{(ij)t}$ standing for the observed scores on the i^{th} item of the j^{th} child.

$$(5) \quad Y_{(ij)t} = b_0 + b_1 * Month_{(ij)t} + [u_{0j} + u_{1j} * Month_{(ij)t} + v_{i0} + e_{(ij)t}]$$

$$(i = 1, 2, \dots, I_j; j = 1, 2, \dots, J).$$

The model consists of two parts: a fixed part and a random part. In the fixed part, the intercept (b_0) and the average change in scores with age (b_1) were estimated. The effect of age (i.e. age_child) on the observed measurements were analysed as both a fixed effect and a

random effect because the data was obtained from children in different age groups and the children in the same age group were not measured at exactly the same age at each testing moment. The factor *age_child* was centred around the mean age of all children (78 months or 6;6, referred to as month zero in our modelling). The age Month of a child at a test moment was represented in the model by dividing the difference between age Month and age zero by 10 in order to avoid extremely small regression weights. The intercept was the average score of the children at month zero. The average score of children of age Month was estimated as $b_0 + b_1 * \text{Month}$. In the random part of the model (between square brackets), four residual terms were distinguished: u_{0j} , u_{1j} , v_{i0} and $e_{(ij)t}$. The residual scores were assumed to be normally distributed with an expected score of 0.0 and a variance of S^2_{u0j} , S^2_{u1j} , S^2_{vi0} and $S^2_{e(ij)t}$ respectively. The variance of the first residual (S^2_{u0j} or S^2_{child}) indicated the variance in scores between children at month zero. The variance of the second residual (S^2_{u1j} or $S^2_{\text{age_child}}$) showed the variance in change in scores with age between children; the change in scores was assumed to be larger for some children than for other children. These two variance components quantified individual differences in production and comprehension. The variance of the third residual (S^2_{vi0} or S^2_{item}) quantified differences in average scores assigned to different items or stimuli. The variance of the fourth residual ($S^2_{e(ij)t}$ or S^2_{error}) indicated the interaction between child and item. This variance component is usually interpreted as an error variance and will not be further discussed.

3.2 Production

Inter-rater reliability analysis on the scores from the three raters showed a high inter-rater agreement: Cronbach's Alpha = .746; Interclass correlation coefficient = .743. The mean production score for each child in each focus condition was computed by averaging the scores from the three raters. The results of multi-level modelling are summarised in

Appendix 3 and illustrated in Appendix 4. In the initial focus condition (left panel of Appendix 4), the estimated average production score was 3.018 at the age of 6;6, the mean age of all participating children. It changed significantly with age ($p = .003$): an increase or decrease of 0.206 at every increase or decrease of 10 months in age relative to the score at the mean age. For example, the average production score was 2.6 at the age of 4;10 but 3.43 at the age of 8;2. Further, the production scores differed significantly between children at the mean age ($S^2_{\text{child}} = .258$, $p < .001$). The change in the production scores with age also differed significantly between children ($S^2_{\text{age_child}} = 0.046$, $p = .004$). Finally, the variance in scores on different items was significant but was relatively small ($S^2_{\text{items}} = 0.025$, $p = .004$).

In the final focus condition (right panel of Appendix 4), the estimated average production score was 3.559 at the mean age. It did not change significantly with age ($p = .247$). Further, production scores differed significantly between children at the mean age ($S^2_{\text{child}} = 0.200$, $p < .001$). But the change in production scores with age did not differ significantly between children ($p = .399$). Finally, the variance in scores on different items was not significant ($p = .399$).

Compared to the production of the adult controls, the children's average production score was significantly lower than that of the adults even at the oldest age (8;8) in the initial focus condition (3.636 in children vs. 3.957 in adults) ($\chi^2 = 14.24$; $df = 1$; $p < .001$). The children's average score would match that of the adults only at the age of 10;3. In the final focus condition, the children's average production score (3.557) was similar to that of the adults (3.586) throughout the age range under investigation ($\chi^2 = 2.13$; $df = 1$; $p = .144$).

In the light of these results, we can conclude that Hypothesis 1a on the children's production is borne out. There was indeed a difference in how target-like the children's production was perceived by adults in different sentence positions. Hypothesis 3 concerning individual differences in production is however largely falsified. Instead of exhibiting stable

individual differences across ages and focus conditions, the children differed significantly from each other in the amount of development with age in initial focus, but not in final focus.

3.3. Results – comprehension

As mentioned in section 2.3, the answers (given by the pets) were segmentally and lexically correct on the experimental trials but contained either lexical or pronunciation errors on the filler trials; the answers were spoken with a contextually inappropriate prosodic pattern on half of the experiment and filler trials. Chen (2010) found that although both adults and children judged the answers on the experimental trials to be mostly correct and the answers on the filler trials to be mostly incorrect, regardless prosody, adults did so even more frequently than children. Post-experiment interviews with children revealed that children rejected the answers in the experimental dialogues as correct answers because of different interpretations of the objects in the pictures or false perception of pronunciation errors. Our modelling on the effect of age and random factors on the children's correct and incorrect judgements on the experimental trials also showed that the average proportion of correct answers increased significantly with age ($p < 0.05$).

We thus decided to model the effect of the fixed factor `age_child` and the variance of the four residuals only on the log-transformed ($\ln$) reaction times on the experimental trials on which the answers were judged to be correct. Extremely short reaction times (< 200 ms) on the selected experimental trial were subsequently excluded from the models. In total, we included 2262 trials (83% of all trials) in the models: 568 trials in the initial focus-appropriate prosody model (85%), 556 trials in the initial focus-inappropriate prosody model (85%), 560 trials in final focus-appropriate prosody model (82%), and 578 trials in the final focus-inappropriate prosody model (80%). To access the difference in reaction times between the appropriate prosody condition and the inappropriate prosody condition, we compared the

estimates of the intercept, the fixed factor `age_child` and the variances of three residual components in these conditions. The results of the modelling and the subsequent Chi-squared tests are summarised in Appendix 5, in which the estimates differing significantly between the two prosody conditions are underscored.

As our hypothesis on comprehension (Hypothesis 1b) concerned the differences in reaction times between the two prosody conditions, we describe the results on the estimates that differed significantly between these two conditions in detail and refer to estimates that did not differ significantly only when necessary. In the initial focus condition, the average (ln) reaction time was significantly longer in the appropriate prosody condition (6.92) than in the inappropriate prosody condition at the mean age (6.887) ($\chi^2 = 6.243$; $df = 1$; $p = .012$). The average (ln) reaction time changed with age to similar degrees in the two prosody conditions ($\chi^2 = 1.611$; $df = 1$; $p = .204$): it decreased/increased by 0.119 at every increase/decrease of 10 month in age relative to the mean age in the appropriate prosody condition and by 0.129 in the inappropriate prosody condition. The average (ln) reaction time thus remained longer in the appropriate prosody condition than in the inappropriate condition over the whole age range tested in our study (i.e. 4;2 – 8;8). Furthermore, the reaction times differed between children to a similar degree at the mean age in the two prosody conditions ($\chi^2 = 0.874$; $df = 1$; $p = .35$). The change in the reaction times with age differed between children to a significantly larger degree in the appropriate prosody condition than in the inappropriate prosody ($\chi^2 = 64.489$; $df = 1$; $p < .001$). Finally, the change in the reaction times on different items was not significant between the two prosody conditions.

In the final focus condition, the average (ln) reaction time was significantly shorter in the appropriate prosody condition (6.699) than in the inappropriate prosody-to-focus mapping condition (6.715) at the mean age ($\chi^2 = 4.308$; $df = 1$; $p = .038$). The average (ln) reaction time changed with age to a significantly larger degree in the appropriate prosody condition

than in the inappropriate prosody condition ($\chi^2 = 4.628$; $df = 1$; $p = .031$): It decreased/increased by 0.152 at every increase/decrease of 10 month in age relative to the mean age in the appropriate prosody condition but by 0.109 in the inappropriate prosody condition. The average (ln) reaction time thus remained shorter in the appropriate prosody condition than in the inappropriate prosody condition in the whole age range but the difference in the average (ln) reaction times between the two prosody conditions became larger at older ages due to a faster decrease in the appropriate prosody condition than in the inappropriate prosody condition. Furthermore, the reaction times differed between children at the mean age to a larger degree in the appropriate prosody condition (0.071) than in the inappropriate prosody condition (0.065) ($\chi^2 = 23.02$; $df = 1$; $p < .001$). The change in the reaction times with age differed between children to a significantly larger degree in the inappropriate prosody condition (0.019) than in the appropriate prosody (0.026) ($\chi^2 = 106.6$; $df = 1$; $p < .001$). Finally, the change in the reaction times on different items was not significant between the two prosody conditions.

These results show that at the group level, the children responded significantly slower in the appropriate prosody condition than in the inappropriate prosody condition in initial focus in the whole age range tested in this study. The opposite pattern was found in final focus. In other words, the children were adult-like in comprehension in final focus from 4;8 onwards but were not adult-like in initial focus even at the age of 8;8. Hypothesis 1b is thus only borne out in the final focus condition. Furthermore, the children not only differed significantly from each other in the reaction times at the mean age in the final focus condition but also differed significantly in the change in reaction times with age in both focus conditions and exhibited larger differences in reaction times between the two prosody conditions at an older age in the final focus condition. Hypothesis 3 on stable individual differences in comprehension can thus be rejected.

3.4 The link between production and comprehension

The results from our production and comprehension experiments show that at the group level the children were adult-like in neither production nor comprehension throughout the age range under investigation in the initial focus condition but were adult-like in both production and comprehension from the youngest age onwards (i.e. 4;8) in the final focus condition. To find out whether children's comprehension scores could predict their production scores in each focus condition, we expanded the model in (5) by including the comprehension score as a second fixed factor in addition to the factor age_child. We operationalised 'comprehension score' in each focus condition as the ratio between the average (ln) reaction time in the appropriate prosody condition (APP) and that in the inappropriate prosody condition (INAPP). Each child thus had two comprehension scores, one for each focus condition. If a child responded faster in the appropriate-prosody condition than in the inappropriate-prosody condition, he should have a comprehension score lower than 1; if a child responded slower in the appropriate-prosody condition than in the inappropriate-prosody condition, he should have a comprehension score higher than 1. Comprehension scores lower than 1 was supposed to indicate adult-like comprehension; comprehension scores higher than 1 non-adult-like comprehension, following Birch and Clifton (1995) and Chen (2010). The expanded version of the model in (5) can be written as the equation in (6) with $Y_{(ij)t}$ representing the production score on the i^{th} item of the j^{th} child.

$$(6) Y_{(ij)t} = b_0 + b_1 * Month_{(ij)t} + b_2 * \frac{APP_{it}}{INAPP_{it}} + [u_{1j} + u_{2j} * Month_{(ij)t} + v_{1i} + e_{1(ij)t}]$$

$$(i = 1, 2, \dots, I; j = 1, 2, \dots, J).$$

Moreover, to take individual differences into account, we modelled the relation between production and comprehension for children with categorically different comprehension abilities separately. We recognised two sub-groups among the children on the basis of their comprehension scores, the 'good' comprehenders and the 'poor'

comprehenders. The ‘good’ comprehenders were the children with a comprehension score smaller than 1 (N = 33 in initial focus, N = 36 in final focus); the ‘poor’ comprehenders were the children with a comprehension score equal to or larger than 1 (N = 32 in initial focus, N = 32 in final focus).

The results of our modelling are summarised in Appendix 6. The variances of the residual components were estimated but not included in Table 4, as they were not relevant to the effect of the fixed factor comprehension score. As can be seen, in the initial focus condition, the comprehension score could predict the production score but in different ways for the two types of comprehenders ($p < .001$ in both cases). For the ‘good’ comprehenders, the lower the comprehension score was, the higher the production scores was ($p < .001$). For the ‘poor’ comprehenders, the higher the comprehension score was, the higher the production score was ($p = .015$). Further, the factor age had a significant effect on the production score such that the production score became higher at an older age ($p < .001$ in the case of the ‘good’ comprehenders and $p = .044$ in the case of the ‘poor’ comprehenders). There was no difference in the effect of AGE in the two types of comprehenders ($\chi^2 = 1.44$; $df = 1$; $p > .230$). This result was in line with the results from the multilevel modelling reported in section 3.2. In contrast, in the final focus condition, neither comprehension score nor age was a good predictor for the production score of the two types of comprehenders. This lack of effect of age_child on the production score was in line with the results from the multilevel modelling reported in section 3.2. It was however unexpected that the comprehension scores were not predictive of the production scores at all in the final focus condition.

Taken together, the children’s comprehension was predictive of their production in initial focus but not in final focus, leading partial support to Hypothesis 2. But comprehension predicted production in different ways for different types of comprehenders in initial focus regardless of age, arguably supporting Hypothesis 3.

4. Discussion

We have addressed three questions essential to the issue of the production-comprehension link in the acquisition of prosodic focus marking from a longitudinal perspective: (1) what is the development like in production and comprehension between 4 and 8 years of age? (2) Is children's comprehension predictive of their production? (3) How do children differ in production, comprehension and the relation between production and comprehension? Our results have revealed a complex picture on production and comprehension in the acquisition of prosodic focus marking.

First of all, we have found that development in production and comprehension at the group level was different in different sentence positions, consistent with Hypothesis 1a but contra Hypothesis 1b. In initial focus, the children's production got better as they got older but did not reach the adult-like performance even at the age of 8;8. Their comprehension was not adult-like at the age of 4;8 and did not improve with age, except that they responded faster when they got older regardless of the appropriateness of the focus-to-prosody mapping. In final focus, the children's production reached the adult-like performance at the age of 4;8 and did not show significant changes with age. Their comprehension was also adult-like from at the age of 4;8, i.e. responding faster when the focus-to-prosody mapping was appropriate than when it was inappropriate. Interestingly, the difference in the reaction times between the two conditions became larger with age due to a faster decrease in the reaction times with age in the appropriate prosody condition than in the inappropriate prosody condition.

The slower development in production and comprehension in initial focus than in final focus can be explained by the higher degree of complexity of how prosody is used to encode focus in sentence-initial position than in sentence-final position. As discussed in section 2, in sentence-initial position, adult speakers usually pronounce the subject noun

regardless of information structure with a falling accent but vary the phonetic realisation of the falling accent to distinguish focus from non-focus; whereas in sentence-final position, adult speakers mostly realise the focal constituent with a falling accent and realise the non-focal constituent with other types of accent or deaccentuation. Phonetic realisation concerns relatively small changes in pitch and duration and is perceptually less salient, compared to accent placement and choice of accent type. It may therefore pose more difficulty for children to perceive the form-function mapping between gradient changes in pitch and duration and focus, and to eventually learn to vary phonetic realisation for focus-marking purposes in their own production and to establish the default representation of the form-function mapping in sentence-initial focus. Earlier work on Dutch-speaking children has provided evidence for the slower development in initial focus in production (Chen 2011, 2009). We have shown for the first time that this is also the case in comprehension.

The finding on changes in reaction times with age between the two prosodic conditions suggests that development in the comprehension of the focus-to-prosody mapping may be characterised by an increased ability to integrate contextually appropriate prosody into lexical and syntactical information to comprehend speech, rather than an increased sensitivity to inappropriate use of prosody. This interpretation does not seem compatible with results from studies of brain responses to different types of prosodic information in adults. Typically, inappropriate use of prosody leads to notable brain responses in electrophysiological experiments. For example, inappropriate accent placement (e.g. accents on given information and deaccenting new information) can evoke an N400 – a negative-going wave that peaks around 400 ms after the onset of the critical word and a late centroparietal positivity, which is associated with discourse reorganisation process as a result of the inappropriate mapping between prosody and information structure (see Baumann and Schumacher, 2011 for a review). Our interpretation is nevertheless in accordance with

findings showing that children aged between 4 and 10 seem to take little notice of prosody in performing various comprehensions tasks. For example, Lahey (1974) found that 4- and 5-year-old English-speaking children were not significantly worse at acting out coordinate sentences and sentences with relative clauses when the sentences were spoken with monotonous prosody than when they were spoken with proper prosody. Bates (1976) reported that English-speaking children's imitation of sentences was disrupted by marked word order but not by pragmatically inappropriate accent placement. Morton and Trehub (2001) found that English-speaking children respond primarily to the lexical content of a sentence in perception of emotions such as happy and sad when prosodic cues conflict with lexical cues at the age of 4, their reliance on the lexical content declines between the age of 5 and 10 and become reliant on prosody only at the age of 10. We propose that children in the multiword-utterances stage initially rely on lexical and syntactical information in language comprehension and gradually establish a network of representations of contextually appropriate and statistically more probable form-function mappings between prosodic cues and meanings. In this process, they begin with responding strongly to prosodic information when it appears in the associated context. Only at a more advanced stage, they take the contextually probable form-function mappings as the default and respond in a negative way to contextually improbable form-function mappings. This development may take place for different form-function mappings at different rates because children establish certain form-function mappings earlier than others (see Chen et al., to appear for a review).

Furthermore, we have found that the children's comprehension scores were predictive of their production scores in initial focus but not in final focus, partially supporting our hypothesis on the comprehension-precedes-production asymmetry. The relation between comprehension and production in initial focus is consonant with the role of comprehension as the prerequisite for production, as claimed in theories of language acquisition. However, the

independence of development in production and comprehension in final focus suggests that comprehension ability as tested in the RT paradigm may not be the precondition for production. That is, children can develop adult-like use of prosody in focus marking in production without processing the focus-to-prosody mapping in language comprehension in the same way as found for adults (Birch and Clifton 1995, Cutler et al. 1997, Chen 2010). The question that arises is then how children acquire the production ability.

We propose that children figure out how to use prosody to mark focus through perception of the form-function mappings between prosodic variations and focus conditions (in different contexts) in others' production. Specifically, children may initially operate on Bolinger's universal and gradually grammaticalise the form-function mapping between prosodic prominence and focus by implementing prosodic prominence in a language-specific manner. The process of grammaticalisation is driven by perception of the association between focus and various forms of prosodic prominence in the speech that children are exposed to (i.e. input). It follows that more salient form-function mappings are likely to be easier to perceive and are therefore acquired earlier than less salient form-function mappings. This has been shown to be indeed the case. For example, Dutch-speaking children's use of accent placement and accent type to express narrow focus in final focus became adult-like earlier than their use of relatively subtle phonetic cues in initial focus. Salience can be also in the nature of the focus condition. For example, English- and German-speaking children can use prosodic prominence in an adult-like way in contrastive focus before they do so in (non-contrastive) narrow focus (see Chen 2015 for a review).

The ability to associate prosodic prominence with focus at the perceptual level is not the same as the ability to capitalise on the focus-to-prosody mapping to process sentence meaning in online language comprehension; the former needs not lead to the latter either, in particular in relatively less experienced language users. For example, Ge, Chen and Yip

(2018) have found that advanced Cantonese learners of English tend to perceive the focal word in a sentence as most prominent-sounding but the speed of their sentence comprehension is not affected the appropriateness of the focus-to-prosody mapping in English. Also, listeners can use different strategies in language comprehension and need not all heavily rely on prosodic cues. There may thus be no principled reasons for a coordinated development between production and comprehension in individual children in the acquisition of prosodic focus-marking. Further research on children's perception of the focus-to-prosody mapping is needed to obtain a clearer understanding of the role of perception in prosodic focus marking in production.

Lastly, we have found individual differences in the children's production, comprehension and more importantly, how comprehension was predictive of production. The children differed between each other in the amount of change with age in initial focus in production and in both focus conditions in comprehension. Changes in reaction times with age were larger in the appropriate prosody condition in initial focus but in the inappropriate prosody condition in final focus. The individual differences were thus not stable across age and focus conditions, different from the patterns of individual differences reported for vocabulary and grammar development (Kidd et al. 2018). Comprehension and production were related in different ways for different types of comprehenders in initial focus. For the children with adult-like comprehension ('good' comprehenders), lower comprehension scores predicted higher production scores. For the children with non-adult-like comprehension ('poor' comprehenders), higher comprehension scores predicted higher production scores. These results would seem to suggest that children's production benefits from comprehension only if they are able to integrate prosodic information with lexical and syntactical information in online language comprehension. In contrast, comprehension was not predictive of production in final focus. Given that the children were already adult-like in both production

and comprehension at the beginning of the testing phase, it remains to be tested whether comprehension can be predictive of production at a younger age in final focus.

Our finding on the individual differences in how production is related to comprehension underscores our postulation that comprehension of the focus-to-prosody mapping as measured in the RT paradigm may not be a prerequisite for using prosody to mark focus in an adult-like manner in children. It also suggests that the ability to process the focus-to-prosody mapping in online language comprehension may not be fully developed in many typically developing children. This would predict that not all adults with normal speaking and hearing comprehend sentences faster when the prosody is contextually appropriate than when the prosody is contextually inappropriate. In a preliminary study on individual differences in online comprehension of prosodic focus marking, Lentz and Chen (2015) indeed found that about half of the 18 participants in their study did not show the desired reaction time differences between appropriate prosody and inappropriate prosody. Studies on how parents' production and comprehension are related to children's production and comprehension are needed to find out whether input plays a role in why some children can develop target-like production but do not benefit from appropriate prosody in comprehension.

5. Conclusions

To conclude, we have found that children's production and comprehension of the focus-to-prosody mapping in Dutch develop at different rates in different sentence positions at the group level. Neither becomes adult-like by the age of 8;8 in sentence-initial position but both are adult-like from the age of 4;8 onward. We can explain the slower development in production and comprehension in initial focus than in final focus by referring to the higher degree of complexity of how prosody is used to encode focus in sentence-initial position than

in sentence-final position. The development in comprehension is characterised with a faster decrease with age in reaction times to sentences with an appropriate focus-to-prosody mapping than in reaction times to sentences with an inappropriate focus-to-prosody mapping. This result suggests an increased ability to integrate contextually appropriate prosody with lexical and syntactical information in online comprehension of sentence meaning.

Furthermore, we have found that children's comprehension is predictive of their production in initial focus but not in final focus. This result suggests that the ability to integrate prosodic information with lexical and syntactical information to comprehend speech may not be the precondition for prosodically realising focus in production and children may acquire the use of prosody in focus marking via perception of the form-function mappings between prosody and focus in other speakers' production. Finally, our study has shown that individual differences changes with age in production and comprehension in prosodic focus marking, different from individual differences in other areas of language development. Notably, comprehension and production are related in different ways for different types of comprehenders in initial focus. Children's production appear to benefit from comprehension only if they are able to integrate prosodic information with lexical and syntactical information in online language comprehension. Comprehension is not predictive of production in final focus. This may be related to the fact that children are already adult-like in both production and comprehension in final focus at the youngest age tested in this study. Studies on younger children are needed to find out whether comprehension can be predictive of production at a younger age in final focus.

References

- Baumann, S., & Schumacher, P. B. (2011). (De-)Accentuation and the Processing of Information Status: Evidence from Event-Related Brain Potentials. *Language and Speech*, 55, 361–381. <http://doi.org/10.1177/0023830911422184>
- Bates, E. (1976). *Language in Context*. Academic Press, New York.
- Birch, S., & Clifton, C.J., (1995). Focus, accent, and argument structure: effects on language comprehension. *Language and Speech*, 38 (4): 365–391.
- Bishop, D. (2003). Test for the reception of grammar. Pearson assessment.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International* 5 (9/10), 341–345.
- Bolinger, D. (1978). Intonation across languages. In J. H. Greenberg (ed.), *Universals of human Language* (Vol. 2): *Phonology* (pp.471-524). Stanford: Stanford University Press.
- Bolinger, D. (1983). Intonation and gesture. *American Speech*, 58: 156-174.
- Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, Pussycat? On talking to babies and animals. *Science* 296 (5572), 1435.
- Chen, A., Esteve Gibert, N., Prieto, P., & Redford (to appear). Development in phrase-level prosody from infancy to late childhood. In C. Gussenhoven & Chen, A. (eds) *The Oxford handbook of language prosody*. Oxford: OUP.
- Chen, A. (2018). Get the focus right across languages: Acquisition of prosodic focus-marking in production. In P. Prieto & Esteve-Gibert, N. (eds) *Prosodic development in first language acquisition* (pp. 295-314). Amsterdam: John Benjamins.
- Chen, A. (2018b). The production, perception and comprehension of prosodic granularity in children and adults. Talk given at the Hanyang International Symposium on Phonetics and Cognitive Sciences of Language (HISoPhCS). Hanyang University, Seoul, Korea, 25-26 May 2018

- Chen, A. (2015). Children's use of intonation in reference and the role of input. In L. Serratrice & S. Allen (eds.) *The acquisition of reference* (pp.83-104). Amsterdam: John Benjamins.
- Chen, A. (2012). Prosodic investigation on information structure. In M. Krifka, & R. Musan (eds.) *The expression of information structure* (pp.251-286). Berlin: Mouton de Gruyter.
- In *The expressions of cognitive categories*. Series editors: W. Klein and S. Levinson.
- Chen, A. (2011). Tuning information packaging: intonational realization of topic and focus in child Dutch. *Journal of Child Language*, 38: 1055-1083.
- Chen, A. (2010). Is there really an asymmetry in the acquisition of the focus-to-accentuation mapping. *Lingua*, 120: 1926-1939.
- Chen A. (2009). The phonetics of sentence-initial topic and focus in adult and child Dutch. In M. Vigário, S. Frota and M. J. Freitas (eds.), *Phonetics and Phonology: Interactions and interrelations* (pp. 91-106). Amsterdam: Benjamins.
- Chen, A., den Os, E., and de Ruiter, J. P. (2007): Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech. *The Linguistic Review*, 24 (2-3): 317-344.
- Clark, E. (1993). *The Lexicon in acquisition*. Cambridge: CUP.
- Clahsen, H. (2008). Behavioral methods for investigating morphological and syntactic processing in children. In: Sekerina, I.A., Ferná' ndez, E.M., Clahsen, H. (Eds.), *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*. (pp. 1-28) John Benjamins, Amsterdam.
- Cruttenden, A. (1985). Intonation comprehension in ten-year-olds. *Journal of Child Language*, 12: 643–661.
- Cutler, A., Dahan, D., van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, 40 (2): 141–201.

- Cutler, A., and Swinney, D. A. (1987). Prosody and the development of comprehension. *Journal of Child Language*, 14: 145–67.
- Dahan, D., Tanenhaus, M. K., and Chambers, C. G. (2002). Accent and reference resolution
- Frank, M.C. et al. (2017) Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44: 677–694.
- Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Hoboken, New Jersey: John Wiley & Sons.
- Gussenhoven, C. (2005). Transcription of Dutch intonation. In: Jun, S. (Ed.), *Prosodic Typology and Transcription: A Unified Approach*. (pp. 118-145). Oxford: OUP.
- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: CUP.
- Hendriks, P. (2005). Asymmetries in the acquisition of contrastive stress. In: Paper presented at the workshop on Contrast, Information Structure and Intonation. Stockholm University, October 28–29, 2005 Retrieved from http://www.let.rug.nl/_hendriks on 21 June 2006.
- Hendriks, P., and Koster, C. (2010). Production/comprehension asymmetries in language acquisition. *Lingua*, 120: 1887-1897.
- Ito, K., and Speer, S. R. (2008): Anticipatory effect of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, 58, 541-573.
- Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, 22, 154–169.
<http://doi.org/10.1016/j.tics.2017.11.006>
- Ladd, D. R. (2006). *Intonational Phonology*. Cambridge: CUP.
- Lahey, M., 1974. Use of prosody and syntactic markers in children’s comprehension of spoken sentences. *Journal of Speech and Hearing Research* 17, 656–668.
- Lambrecht, K. (1994). *Information structure and sentence form: Topics, focus, and the representations of discourse referents*. Cambridge: CUP.

- Lentz, T. and Chen, A. (2015). *Unbalanced adult production and perception in prosody*. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow.
- Quené, H. and Bergh, H. van den (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59: 413-425.
- Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development*, 72(3), 834–843. <http://doi.org/10.1111/1467-8624.00318>
- Müller, A., Höhle, B., Schmitz, M., and Weissenborn, J. (2006). Focus-to-stress Alignment in 4- to 5-year-old German-learning Children. In: Belletti, A., Bennati, E., Chesi, C., Di Domenico, E., Ferrari, I. (Eds.), *Proceedings of GALA 2005*. (pp. 393-407). Cambridge Scholars Publishing, Cambridge.
- Roberts, C. (1996). Information structure: Towards an integrated theory of formal pragmatics. In J. H. Yoon & A. Kathol (Eds.), *OSU Working Papers in Linguistics*, Vol. 49. Columbus: The Ohio State University Department of Linguistics.
- Szendroöi, K. (2004). Acquisition evidence for an interface theory of focus. In: van Kampen, J., Baauw, S. (Eds.), *Proceedings of GALA 2003*.
- Snijders, T., and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and applied multilevel analysis*. Sage: London.
- Vallduví, E., and Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, 34(3): 459-520.
- Veenker, T. J. G. (2013). The Zep experiment control application (version 1.6) [Computer software]. Utrecht Institute of Linguistics OTS, Utrecht University. Available from <http://www.hum.uu.nl/uilots/lab/zep>.
- Watson, D., Gunlogson, C., and Tanenhaus, M. (2008). Interpreting pitch accents in on-line comprehension: H* vs L+H*. *Cognitive Science*, 32: 1232-1244.

- Weber, A., Braun, B., and Crocker, M. (2006). Finding referents in time: eye-tracking evidence for the role of contrastive accents. *Language and Speech*, 49(3): 367-392.
- Wells, B., Peppé, S., and Goulandris, N. (2004). Intonation development from five to thirteen. *Journal of Child Language*, 31: 749–78.
- Xu, Y. (1999). Effect of Tone and Focus on the Formation and Alignment of f0 contours. *Journal of Phonetics*, 27: 55–107.
- Xu, Y., & Xu, C.X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, 33: 159–197.

Appendix 1

An example trial of final focus in the picture-matching game

Experimenter (showing a picture to the participant): Look! The girl. It seems that the girl is drawing something. What is the girl drawing?

Participant: (taking a picture from his set and inspecting it) The girl is drawing the spoon.

The experimenter looks for the picture of the spoon in the box and handed over both pictures to the participant.

Appendix 2

The left panel consisting of three drawings is a screenshot of what participants can see on the computer screen on the trials involving the sentences *Het/een varken wast een/de bloes* ‘The/a pig is washing a/the blouse’. The right panel consisting of four figures illustrates the prosodic patterns of the sentences uttered as answers to the question *Wie wast de bloes?* ‘Who is washing the blouse’ (initial focus) and *Wat wast het varken?* ‘what is the pig washing?’ (final focus) in both the appropriate prosody condition and the inappropriate prosody condition.

Note: The examples were transcribed following Gussenhoven (2005). H*L (pronounced as high-star-low) stands for a falling accent; !H*L (pronounced as downstepped high-start-low) a downstepped falling accent, i.e. a falling accent with a lower peak than the preceding accent. H stands for high pitch level and L stands for low pitch level; the ‘*’ sign indicates which pitch level is realised in the stressed syllable. %L (pronounced as percent-L) stands for an initial low boundary tone and L% (pronounced as L-percent) a final low boundary tone.

Appendix 3

Parameter estimates for the production scores in the initial and final focus conditions.

	Initial focus			Final focus		
	Estimate	(se)	p	Estimate	(se)	p
Fixed Parameters						
Intercept (b_0)	3.016	-0.108	<.001	3.559	-0.067	<.001
Age_child (b_1)	0.206	-0.065	0.003	0.045	-0.046	0.247
Random Parameters						
S^2_{child}	0.258	-0.031	<.001	0.200	-0.017	<.001
$S^2_{\text{age_child}}$	0.046	-0.015	0.004	0.000	0.000	0.399
S^2_{items}	0.025	-0.008	0.004	0.000	0.000	0.399

Note: The factor age_child was centred around the mean age of all children (i.e. 6;6 or 78 months). A unit of change in age was defined as an increase or decrease of 10 months relative to the mean age.

Appendix 4

Change in estimated production scores (y-axis) with age (x-axis) on average (M: dark line) and for individual children in the initial-focus (left) and final-focus (right) conditions. Each gray line represents a child at two testing moments, with the left end of the line indicating the production score at the first testing moment and the right end of the line indicating the production score at the second testing moment. The letter 'A' indicates the average production score of the adults, aligned with the point of 108 months for comparison reasons.

Appendix 5

Parameter estimates for reaction times in the initial-focus condition and the final-focus condition.

	Initial focus						Final focus					
	Appropriate prosody			Inappropriate prosody			Appropriate prosody			Inappropriate prosody		
	Est.	(se)	p	Est.	(se)	p	Est.	(se)	p	Est.	(se)	p
	Fixed parameters			Fixed parameters			Fixed parameters			Fixed parameters		
Intercept (b_0)	<u>6.92</u>	-0.041	<.001	<u>6.887</u>	-0.039	<.001	<u>6.699</u>	-0.045	<.001	<u>6.715</u>	-0.044	<.001
Age_child (b_1)	-0.119	-0.031	<.001	-0.129	-0.028	<.001	<u>-0.152</u>	-0.034	<.001	<u>-0.109</u>	-0.034	0.030
	Random parameters			Random parameters			Random parameters			Random parameters		
S^2_{child}	0.054	-0.002	<.001	0.046	-0.002	<.001	<u>0.071</u>	-0.002	<.001	<u>0.065</u>	-0.022	<.001
$S^2_{age_child}$	<u>0.018</u>	-0.001	<.001	<u>0.009</u>	-0.001	<.001	<u>0.019</u>	-0.001	<.001	<u>0.026</u>	-0.013	0.054
S^2_{items}	0.002	0.000	0.053	0.036	-0.004	<.001	0.001	0.000	0.021	0.000	0.000	0.261

Note 1: The factor age_child was centred around the mean age of all children (i.e. 6;6 or 78 months). A unit of change in age was defined as an increase or decrease of 10 months relative to the mean age.

Note 2: The estimates differing significantly between the two prosody conditions were underscored.

Appendix 6

Parameter estimates for the relation between production and comprehension in the initial and final focus conditions.

	Initial focus			Final focus		
	Estimate	(se)	p	Estimate	(se)	p
	'good' comprehenders			'good' comprehenders		
Intercept	3.288	-0.210	< .001	3.400	-0.140	< .001
Age_child	0.370	-0.080	< .001	-0.810	-0.690	0.200
Comprehension	-12.960	-2.510	< .001	0.120	-0.080	0.130
	'poor' comprehenders			'poor' comprehenders		
	Estimate	(se)	p	Estimate	(se)	p
	'good' comprehenders			'good' comprehenders		
Intercept	2.970	-0.200	< .001	3.690	-0.150	< .001
Age_child	0.230	-0.110	0.044	-0.710	-0.580	0.189
Comprehension	6.350	-2.470	0.015	0.020	-0.090	0.389