Spontaneous speech intelligibility : early cochlear implanted children versus their normally hearing peers at seven years of age

Spontaneous speech intelligibility: early cochlear implanted children versus their normally hearing peers at seven years of age

(Received 24/02/20, Revised 28/01/21, Accepted 28/09/21)

Authors

Nathalie BOONEN (corresponding author)

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: n.boonen@fontys.nl

Hanne KLOOTS

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: hanne.kloots@uantwerpen.be

Pietro NURZIA

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: pietro.nurzia@uantwerpen.be

Steven GILLIS

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

E-mail: steven.gillis@uantwerpen.be

Acknowledgements: We especially thank Jolien Faes for preparing and running the Qualtrics rating scale experiment. Thanks are also due to the action editor and two anonymous reviewers for the constructive comments. This project was funded by a predoctoral research grant of the Research Foundation – Flanders (FWO) to the first author (1100316N). This study was approved by the Ethics Committee for the Social Sciences and Humanities (SHW_15_37) of the University of Antwerp. The authors declare no conflict of interest.

Keywords: Intelligibility; Spontaneous speech; Children with a cochlear implant

Abstract

Speaking intelligibly is an important achievement in children's language development. How far do congenitally severe-to-profound hearing-impaired children who received a cochlear implant (CI) in the first two years of their life advance on the path to intelligibility in comparison to children with typical hearing (NH)?

Spontaneous speech samples of children with CI and children with NH were orthographically transcribed by naïve transcribers. The entropy of the transcriptions was computed to analyze

their degree of uniformity. The same samples were also rated on a continuous rating scale by another group of adult listeners.

The transcriptions of the NH children's speech were more uniform, i.e., had significantly lower entropy, than those of the CI children, suggesting that the latter group displayed lower intelligibility. This was confirmed by the ratings on the continuous scale. Despite the relatively restricted age ranges, older children reached better intelligibility scores in both groups.

Introduction

Reaching intelligible speech is an important milestone in children's speech and language development. For children with a severe-to-profound hearing impairment who received a cochlear implant, becoming as intelligible as their normally hearing peers is an ultimate goal of their rehabilitation. Intelligibility is often viewed as a crucial benchmark because it "requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered" (Freeman et al., 2017: 278). A child who is intelligible for the outside world, can be considered to have acquired and developed these crucial components. As such, intelligibility is considered to be the most practical single index to apply in assessing competence in oral communication (Kent et al., 1994; Subtelny, 1977: 183). Consequently, measures of speech intelligibility are often applied as diagnostics for speech therapy. According to Gordon-Brannan and Hodson (2000), when one third of the continuous speech of a four-year-old cannot be transcribed correctly by others, this child is a candidate for speech therapy. Because of the general importance of intelligibility and because intelligibility scores can give an

indication of whether or not speech therapy is advisable for particular children, speech intelligibility measures are considered “the gold standard for assessing the benefit of cochlear implantation” (Chin et al., 2012: 356).

In the present study, intelligibility is conceptualised as the extent to which the elements (i.c., words) in an acoustic signal generated by a speaker, can be correctly recovered by a listener (Freeman et al., 2017; van Heuven, 2008; Whitehill & Ciocca, 2000). For instance, in a transcription task, intelligibility refers to the extent to which a transcriber can identify the words contained in an utterance. For typically developing children, speech is estimated to be intelligible for all listeners, including those not familiar with the child, around the age of four (Baudonck et al., 2009; Bowen, 2011; Chin & Tsai, 2001; Chin et al., 2003; Flipsen, 2006; Weiss, 1982). For instance, Flipsen (2006) compared the intelligibility of children's conversational speech between the ages of 3;01 and 8;05 using different measures. He found that, irrespective of the specific measure used in the analysis, children were already highly intelligible between 4;0 and 5;0, with scores ranging from 88% to 100%. More recently, Hustad et al. (2020) studied the mean percentage of intelligible words in normally hearing (NH) children’s imitated speech between 2;06 and 3;11. They found a steady increase of the mean intelligibility of multiword utterances from 40% at 2;06, 55% at 3;0, 66% at 3;06 and 78% at 3;11. This means that approximately three out of four words of a four-year-old can be identified by an adult listener not familiar with the child. Thus, the literature on NH children shows that their intelligibility increases with chronological age. Older children tend to be more intelligible than younger ones. However, this does not mean that even 10-year-olds are fully intelligible (Grandon et al., 2020).

In the current study, the speech intelligibility of children with a cochlear implant (CI) is investigated in comparison with that of peers with normal hearing. A CI partially restores a severe-to-profound sensorineural hearing loss. Even though the signal provided by a CI is still degraded compared to the signal in normal hearing (Drennan & Rubinstein, 2008), the device enables children with severe-to-profound hearing impairment to perceive speech and other environmental sounds. After cochlear implantation, children's speech perception has been shown to improve considerably and as a result cochlear implantation is also beneficial for speech and language production (O'Donoghue, 2013; Wie et al., 2020). Research has shown that children with CI can attain spoken language skills similar to those of their normal hearing peers after three to four years of device use (i.a. Bruijnzeel et al., 2016; Dettman et al., 2016; Geers & Nicholas 2013; Wie et al., 2020). However, the population of children with CI is characterized by remarkable variation. On the one hand, variation relates to differences between individual children: while a considerable number of children with CI appear to catch up with their NH peers, some do not catch up at all (Nicholas & Geers 2007; Geers et al., 2016; Duchesne & Marschark, 2019). On the other hand, variation also relates to differences between domains: some areas of speech and language appear to be more difficult to master than others (Duchesne et al., 2019). For instance, Faes et al. (2015) showed that in a group of children with CI acquiring Dutch, inflectional morphology and sentence length (as a proxy of syntagmatic development) were age-appropriate when the children were 7;0, but the former (and not the latter) was already age-appropriate at age 5;0. Moreover, the phonetics of the same children's production of vowels was still significantly different from the vowels of their NH peers at the age of 7;0 (Verhoeven et al., 2016). Thus, although children with CI start with an initial delay in spoken language, a quite significant group eventually reaches age appropriate levels of linguistic functioning. But the

individual variation is also quite large: while some children do catch up with their normally hearing peers, others do not achieve much language comprehension and production even after five years of device use (Barnard et al., 2015).

As to intelligibility, most studies found that CI children's speech intelligibility is less well developed than that of their NH peers (i.a. Castellanos et al., 2014; Chin & Kuhns, 2014; Freeman et al., 2017; Grandon et al., 2020). For instance, Freeman et al. (2017) compared the intelligibility of 24 children with CI, mean age 4;02, with on average almost three years of device use, with 30 NH age-matched peers. On the BIT test (Osberger et al., 1994) in which children are asked to imitate short utterances, the children with CI reached an intelligibility score of 51% (range 0.8% - 95.5%) and the children with NH a score of 84% (range 52.1% - 99.3%). On a retest one year later, both groups' intelligibility score had increased to 67.7% (range 6.1%-98%) for the children with CI and 90.4% (range 78.9%-95.6%) for the children with NH. Even at the age of 9;05 and with on average seven to eight years of device use, the children with CI's intelligibility remains significantly lower than that of children with NH (Chin & Kuhns, 2014). Thus it can safely be concluded that, in general, children with CI are less intelligible than their NH peers, and that there is more individual variation in the intelligibility of children with CI than in NH children.

What causes the variation of children with CI's speech and language development and their intelligibility in addition to the variation which can be expected from children with NH hearing? This issue is still high on the research agenda (i.a. Houston et al., 2012; Duchesne et al., 2019; Bavin et al., 2018). Many factors have been shown to contribute to the success of spoken

language development of children with CI, including (1) audiology related factors, such as the age at implantation, the duration of device use, bilateral (or contralateral) cochlear implantation and the children's preoperative and postoperative hearing levels. (2) Child related factors, such as the cause of the hearing impairment (genetic, infections), gender, additional disabilities (mental retardation, speech motor problems), and (3) environmental factors, such as communication modality. An overview is provided in (Boons et al., 2012; Fagan et al., 2020; Gillis, 2018; Niparko et al., 2010). A factor of particular importance here is age. Studies have shown that chronological age is an important factor for intelligibility: as they grow older, children's intelligibility increases irrespective of their hearing status (Grandon et al., 2020). But in the case of children with CI, age is a complicated factor, since it can not only refer to children's chronological age (as is the case for children with NH), but also to the children's so-called hearing age, which is the amount of time between the activation of their device and their chronological age. For instance, a child implanted at the age of 1;0 has a hearing age of two years at the age of 3;0. In addition, the age at implantation has been shown to play a critical role in children's spoken language achievements. In general, earlier implantation appears to lead to better results than later implantation in several domains (Boons et al., 2012; Niparko et al., 2010). But the research findings with respect to the effect of the variable age on children with CI's intelligibility are not unequivocal. In some studies, a significant effect of chronological age on children's intelligibility was found (i.a. Habib et al., 2010; Flipsen & Colvard 2006; Grandon et al., 2020) but not in others (e.g., Khwaileh & Flipsen 2010). Hearing age was found to be a significant predictor of intelligibility by i.a. Flipsen and Colvard (2006), but hearing age was not always considered as a predictor. Age at implantation predicted children's intelligibility in a considerable number of studies (i.a. Habib et al., 2010; Svirsky et al., 2007; Montag et al., 2014;

Grandon et al., 2020) but this was not the case in other studies (i.a. Flipsen & Colvard 2006; Khwaileh & Flipsen 2010). Nevertheless, a general finding appears to be that earlier implantation leads to better results in speech and language development and in intelligibility. At present there is consistent evidence that implantation in the first two years of life leads to consistently better results in spoken language development in comparison to later implantation, and even (inconclusive) evidence for even better outcomes of implantation in the first year of life (Bruijnzeel et al., 2016; Dettman et al., 2016).

In the present study, the intelligibility of congenitally hearing-impaired children with a cochlear implant was assessed in comparison with that of normally hearing seven-year-old peers. The children were implanted on average around their first birthday, and all demographic variables were held constant as far as possible (see Method section).

Measuring intelligibility

In the studies reviewed so far, children's speech intelligibility was assessed in many different ways. The methods can be situated relative to two dimensions: (1) the amount of control that the investigator exerts on the material that is collected and analyzed; and (2) the analytic versus holistic nature of the assessment method, or subjective ratings versus objective ratings (Hustad et al., 2020). With respect to the first dimension, the vast majority of studies used read or imitated speech (i.a. Castellanos et al., 2014; Chin et al., 2012; Chin & Kuhns, 2014; Freeman et al., 2017; Khwaileh & Flipsen, 2010; Montag et al., 2014). Using imitated or read aloud speech has several advantages over spontaneously produced speech. For instance, an examiner has a large amount of control over the stimuli so that it is easy to compare a target word or utterance with the child's production. This makes it straightforward to quantify the overlap between the child's rendition and the target. This controlled approach can be useful for speech and language pathologists who use the results of the intelligibility test as a starting point for their child-specific speech therapy (Flipsen, 2006). However, read or imitated speech have been suggested to be "rather poor predictors of scores for connected speech and everyday performance with hearing aids" (Cox & McDaniel, 1989: 347), especially for clinical populations such as hearing-impaired children (Ertmer, 2010).

Spontaneous speech is an alternative for read or imitated speech in assessing speech intelligibility. The most important advantage of spontaneous speech is its greater ecological validity. In other words, spontaneous speech is more comparable to everyday informal speech. Despite this major advantage, only few studies use spontaneous speech for assessing children's

speech intelligibility (i.a. De Raeve, 2010; Lejeune & Demanez, 2006; Tye-Murray et al., 1995; Van Lierde et al., 2005). This may be due to the lack of control over the speech sample: whereas in read or imitated speech, the investigator or the clinician decides on the words or utterances that the child is asked to read or imitate, this control is far less in spontaneous speech because the child decides what to say. Hence, in computing the degree of intelligibility, a straightforward measure such as the number or percentage of words read or imitated correctly cannot be relied on, since there is no predetermined set of words or sentences to be produced. This calls for a measure that does not rely on checking if what the child produced equals what the child was supposed to produce. In the present paper such a method will be proposed.

As to the second dimension, measures of the intelligibility can be categorized as "subjective ratings" versus "objective ratings" (Hustad et al., 2020). Subjective ratings use a continuous or an ordinal rating scale on which a holistic, personal perception of a speaker's intelligibility is represented. Probably the most frequently used rating scale is the Speech Intelligibility Rating (SIR) developed by Cox and McDaniel (1989) (i.a. Calmels et al., 2004; De Raeve, 2010; Flipsen, 2008; Lejeune & Demanez, 2006; Toe & Paatsch, 2013). The SIR requires that participants score a child's speech on a five-point scale with a verbal description for each score, ranging from *unintelligible speech even for an adult familiar with the child* to *completely intelligible for all listeners*. Rating scales such as the SIR offer a valid indication of the children's speech intelligibility (AlSanosi & Hassan, 2014; Fang et al., 2014; Flipsen, 2008), especially for assessing the intelligibility of very young children or children with CI implanted at a relatively late age, e.g., late kindergarten (Baudonck et al., 2010; De Raeve, 2010; Toe & Paatsch, 2013). The reason is that children soon reach ceiling scores on the SIR. For instance, De

Raeve (2010) investigated the intelligibility of children implanted before 18 months of age, and found that three years after implantation, the 50th percentile of the group of 45 children scored at the highest level of the SIR. This ceiling score indicates that – according to the SIR – their speech is intelligible to all listeners. However, it is not clear, for instance, whether intelligibility for all listeners pertains to all of the children's speech or only to a limited or particular portion. In other words: children may be considered to be very intelligible according to rating scales but there may still be unintelligible parts in their speech (Miller, 2013). Or children may be rated as "completely intelligible" on the SIR rating scale, but one child may still be more intelligible than another, a difference that cannot be captured using SIR. In this respect, a continuous rating scale, such as the one used in the present study may offer a more diversified picture of children's intelligibility.

"Objective ratings" or analytic ratings take a different approach towards measuring speech intelligibility. Typically, listeners phonetically or orthographically transcribe children's speech. In the case of the imitated or read aloud speech, calculating intelligibility then amounts to applying some measure of overlap between the intended targets and the transcription of the listener(s). But calculating the intelligibility score based on a transcription is not straightforward because a clear target is missing (Flipsen, 2006; Flipsen & Colvard, 2006; Lagerberg et al., 2014). Alternative methods have been proposed that rely on the number of (un)intelligible syllables or words, but these are not unproblematic neither (Flipsen, 2006; Lagerberg et al., 2014; Strömbergsson et al., 2020).

Since transcriptions of spontaneous speech are difficult to judge in terms of correct or incorrect, the method explored in the present study abandons this dichotomous choice and instead makes use of multiple transcriptions. The intelligibility of the speech material is quantified relative the entropy of the transcriptions. Entropy was originally developed in information theory (Shannon, 1948) as a measure that expresses the degree of disorder ("chaos") in data. In linguistic research, entropy measurements were already used for investigating the mutual intelligibility of two closely related languages such as Swedish and Danish (Frinsel et al., 2015; Moberg et al., 2007). In the present context, the assumption is that if a child is highly intelligible, the transcriptions of several listeners will show much uniformity, the degree of disorder or chaos will be low, and, hence, the entropy will be low. Alternatively, if the child's speech exhibits lower intelligibility, the transcriptions will be less uniform, more chaotic, and will thus have a higher entropy score.

Aims of this study

The aim of the present study was to investigate the intelligibility of primary school aged NH and CI children's spontaneous speech. The children were all approximately seven years old, and the children with CI received their device on average at 1;0, and at the time of testing had minimally five years of device experience. The entropy of multiple transcriptions of the children's utterances was used as an index of their intelligibility. It was expected that the children with CI produced speech which was at best as intelligible as the speech of their NH peers. However, given the fact that the NH children had at least one more year of hearing experience, this could be the cause for a lasting advantage of the NH children's intelligibility. A second expectation related to the extent of variability in the two groups of children. Following the reported trends in

the CI literature, it was expected that the entropy scores would show greater variability between subjects with CI than between subjects with NH (Castellanos et al., 2014; Freeman et al., 2017; Montag et al., 2014; Nittrouer et al., 2014; Peng et al., 2004; Yanbay et al., 2014; Young & Killen, 2002). Therefore, the analysis will proceed in two steps. First of all, the intelligibility of children with CI and NH will be compared at a group level. Secondly, the individual variation between the children will be investigated, and the specific demographic variables pertaining to the children with CI will be examined.

A secondary aim of the present study was to examine the relation between the entropy scores obtained from the transcription task and the scores obtained from the holistic judgements on a continuous rating scale. It was assumed that the entropy of the transcriptions was an index of the intelligibility of the children. If this assumption was correct, the entropy scores derived from a comparison of different transcriptions were expected to show some degree of correlation with other measures of speech intelligibility, such as the score on a rating scale. In other words, we expected a correlation between the entropy scores resulting from the "objective" measurement of entropy, and the "subjective" measure of raters' judgements of intelligibility[1].

[1] Thanks are due to an anonymous reviewer of JCL for pointing this out.

Method

The aim was to assess the intelligibility of the spontaneous speech of children with CI and children with NH. An experiment was set up in which speech samples were used which originated from children's spontaneous speech. The participating children and the method of collecting and selecting appropriate stimuli for the experiment will be described first. In the experiment, the children's speech was transcribed by a group of listeners and the same samples were rated by another group of listeners. The participants in the two tasks and the experimental procedure will be described. Finally, the processing of the data resulting from the two experimental tasks and the statistical analyses will be elaborated on.

Stimuli: participating children

In this study, spontaneous speech samples of NH children and children with CI were judged. The parents of the NH and the CI group belonged to the mid-to-high SES stratum as estimated by the Hollingshead Index (Hollingshead, 1975), were native speakers of Belgian Dutch living in Flanders. The control group consisted of sixteen children with NH (ten girls, six boys), native speakers of Belgian Dutch. They were enrolled in the mainstream education system and had no reported hearing loss or additional disabilities as could be judged from the outcome of the UNHS and parental report. At the time of the recording, these children were on average 7;2 years old (SD = 0;7). Their chronological age was comparable to that of the children with CI (Wilcoxon rank sum test: $z = –0.11382$, $p = 0.9094$).

Sixteen children with CI (ten boys, six girls) participated in this study. They were all native speakers of Belgian Dutch, living in Flanders, the Dutch speaking area of Belgium. Their parents were native speakers of Dutch with no self-reported hearing impairment, raising their children orally (monolingual Dutch) with a limited support of signs. The children's hearing impairment was established by the Universal Neonatal Hearing Screening (UNHS) using automated Auditory Brainstem Response hearing tests for newborns, which was administered as a standard procedure in the first weeks of life in Flanders. After the identification of their hearing loss, the children were referred to a specialized audiological centre for further audiological workup. They received acoustic hearing aids and their progress was further monitored. Since their auditory progress was deemed insufficient, they were enrolled as candidates for cochlear implantation. CI candidacy included bilateral hearing loss of at least 85dBHL (up to 2019). All children were implanted before the age of two (mean = 1;0 (years;months), SD = 0;5). Eleven children underwent sequential bilateral implantation, two of them were simultaneously implanted bilaterally. At the time of the recording, the children were between six and eight years old (mean = 7;02, SD = 0;09), and had a minimum of five years of device use, with an average of 6;02 (SD = 0;10). Prior to implantation, their average pure tone average (PTA) was 114 dB HL (SD = 9 dB HL). Their average aided hearing threshold was 29 dB HL (SD = 9 dB HL). Detailed information on the individual children is provided in Table 1. Their medical records and the treating audiological center did not mention any other additional health or developmental issues. Hence, there were no known additional comorbidities beside their hearing-impairment. At the time of the recording, all the children were enrolled in the mainstream education system.

Child	Gender	Etiology	Age at implantation (years; months)	Age at recording (years; months)	Length of device use (years; months)	PTA unaided (dB HL)	PTA aided (dB HL)	Implant type	Speech processor
CI 1	femal e	Genetic	1;2 (6;3)	7;1	5;11	120	35	Nucleus 24 & Freedom	Nucleus Freedom
CI 2	femal e	CMV	0;10 (5;10)	7;1	6;3	115	25	Nucleus 24 & Freedom	Nucleus Freedom
CI 3	male	Genetic	1;6	7;1	5;7	113	42	Nucleus 24	Nucleus Freedom
CI 4	male	Genetic	1;5 (6;4)	7;1	5;8	93	32	Nucleus 24 & Freedom	Nucleus Freedom

CI 5	male	Genetic	0;9	7;2	6;5	120	37	Nucleus 24	Nucleus Freedom
CI 6	femal e	Genetic	0;5 (1;3)	7;1	6;8	117	17	Nucleus 24	Nucleus Freedom
CI 7	femal e	Unkno wn	1;7	7;0	5;5	112	42	Nucleus 24	Nucleus Freedom
CI 8	femal e	Genetic	0;7 (0;7)	5;8	5;0	120	19	Nucleus Freedom	Nucleus 6
CI 9	male	CMV	0;10 (1;8)	8;8	7;10	120	33	Nucleus Freedom	Nucleus 5
CI 10	femal e	Unkno wn	0;10 (1;11)	6;11	6;1	120	20	Nucleus Freedom	Nucleus 6
CI 11	male	CMV	1;7 (1:7)	7;1	5;6	120	15	AB HiRes 90K	Naída CI Q70
CI 12	male	CMV	0;7 (2;2)	6;4	5;9	106	23	Nucleus Freedom	Nucleus 6

CI 13	male	CMV	1;7 (7;3)	7;9	6;2	120	35	Nucleus Freedom & Profile	Nucleus 5&6
CI 14	male	CMV	0;10 (1;9)	7;9	6;11	114	27	Nucleus Freedom	Nucleus 6
CI 15	male	Genetic	0;9 (2;10)	6;8	5;11	114	35	Nucleus Freedom	Nucleus 6
CI 16	male	Genetic	0;11 (2;8)	8;8	7;9	95	27	Nucleus Freedom	Nucleus 6

Table 1: Characteristics of the CI children: their gender (male/female), etiology of their hearing impairment (genetic, CMV infection, unknown), age at implantation and between brackets the age at the second implant, their age and length of device use at the moment of recording, their aided and unaided hearing thresholds (dB HL = decibels hearing level), their implant type and speech processor.

Stimuli: Recording and selection

Audio recordings were made of the children in a quiet room in the comfort of their home or school. The children were asked to tell a story cued by the picture book "Frog, where are you" (Mayer, 1969). Before starting the recordings, the children were allowed to flip through the booklet and look at the pictures. Next, they were asked to tell the story to the researcher and/or caregiver who "did not know the story". The children were stimulated to tell the story independently, but if needed the caregiver or the researcher encouraged and helped the child.

The recordings were orthographically transcribed with the CLAN editor in CHAT format (MacWhinney, 2000). The transcriptions were only used in the selection process of the stimuli for the experiment. In the first step, all the utterances of approximately seven words were selected (e.g. Dutch: "De jongen is bang van de uil", English: "The boy is afraid of the owl"). Then, the corresponding audio fragments were checked. Fragments with background noise, crosstalk and the like were not retained. In addition, utterances with long hesitations, revisions or non-words were also excluded, as well as syntactically ill-formed or incomplete sentences. Finally, a selection of ten utterances was randomly made for each child with NH and each child with CI, resulting in a total of 320 stimuli for the experiment.

The 320 stimuli were divided into five series of 64 utterances. Each series contained two utterances of each CI and NH child, which were randomly selected (without replacement) from the final selection of 10 utterances per child. These five series of 64 utterances were entered into the online tool Qualtrics (Qualtrics, 2005).

Procedure

The experiment consisted of two tasks: a transcription task and a rating task. Two different and non-overlapping groups of participants were recruited for the tasks in which the same series of stimuli were used.

Transcription task

One hundred language students at the University of Antwerp participated in the transcription study. They were native speakers of Belgian Dutch without self-reported hearing problems and without any particular experience with the speech of hearing-impaired children. They were on average 23 years old (SD = 5). The experiment was performed on campus in a computer lab. The students sat in front of a computer screen with headphones which they could set at a comfortable level. The participants were divided into five groups. Each group of 20 students was assigned one of the five Qualtrics series and transcribed all 64 stimuli of that series, resulting in 20 transcriptions of each utterance. Each stimulus could be repeated only three times.

Prior to the actual experiment the participants were instructed on how to transcribe. Examples were given in order to ensure that the instructions were correctly understood. More precisely, the listeners were instructed to use only existing Dutch words in standard orthography and to represent the utterances as accurately as possible. This implied that they should not correct the linguistic errors which are typical for children's speech, such as errors against grammatical gender, the use of erroneous verb declinations, etc. For unintelligible speech, the symbol 'X' was

the agreed upon transcription symbol. In other words, the listeners were instructed to write one X to replace an unintelligible word, an unintelligible part of an utterance or a completely unintelligible utterance.

Rating task

One hundred and fifty students enrolled in the applied linguistics program at the University of Antwerp participated in the rating task. They were all native speakers of Belgian Dutch without self-reported hearing problems and without any particular experience with the speech of hearing-impaired children. They were an average 20 years old (SD = 4). The students completed the rating task at home on their own computer. They were instructed to use headphones to complete the task but received minimal further instructions. On entering the online tool Qualtrics, they saw the instruction: "Duid aan door te klikken of te slepen hoe verstaanbaar deze zin was op een schaal van 'zeer onverstaanbaar' tot 'zeer verstaanbaar'" (Eng.: *indicate by clicking or dragging the slider how intelligible the sentence is on a scale from "fully unintelligible" to "fully intelligible"*). Underneath that instruction the slider represented in Figure 1 was shown together with a play button and a proceed button. Each stimulus could be repeated only three times. The initial position of the slider was always at the far left of the scale, and only the middle point of the scale was indicated by three vertical dashes.

Figure 1: Representation of the Qualtrics screen in the rating experiment.

The experiment was presented to the students as a listening exercise, and they had to use their listening experience to write a short essay on the topic “what is intelligible speech?” as part of their course credit.

Data analysis

Transcription task

Processing the data of the transcription task proceeded in two steps: (1) aligning the transcriptions of the participants of each sentence and (2) computing the entropy of the aligned transcriptions.

Transcription task: Alignment of the transcriptions

The transcriptions of the participants were aligned at the word level. This procedure was repeated for each stimulus separately. As an example, five transcriptions of the same stimulus are provided in Table 2, together with a literal English translation. It can readily be seen in Table 2 that the first transcription (the row indicated by *Transcription participant 1*) contains five words: "de jongen ziet de kikker". The transcription of the second transcriber contains only four words and the transcriber used the symbol X to indicate that the last word was unintelligible. Thus, aligning the transcriptions amounts to the following: the transcribers wrote in a free text field (in Qualtrics) and the 20 transcriptions of each utterance needed to end up in a column like grid structure like Table 2.

A first version of the alignment was automatically produced by a Python script, the output of which was manually checked and adjusted – if needed – in order to maximally align words appropriately. The principal task of the script was to find (nearly) matching words in the orthographic transcriptions and aligning them (see e.g., the five instances of *de* 'the' in the column $Word_1$ of Table2 or the four instances of *jongen* 'boy' in the next column of Table 2). If there was no exact match of the words (e.g., *hond* 'dog' in the transcription of participant 5 in Table 2), the alignment took into account the length of the transcriptions and a word's position. If the transcription length matched, (non-identical) words were aligned if they were on the same position (e.g., the word *jongen* 'boy' in the transcription of the first four participants was aligned with the word *hond* 'dog' of the fifth participant). If the transcription length did not match, the script looked further along the utterance and left blank spaces (indicated as "-----" in the transcription of participant 3 in Table 2) until finding (nearly) matching words (see *kokkin* 'cook' in the transcription of participant 3 in Table 2 which nearly matches *kikker* 'frog' and *kikkers* 'frogs' of participant 1 and 4).

Transcription task: Computing entropy

Given the aligned transcriptions, their relative entropy was calculated using Equation 1. This formula is based on Shannon's original formula of entropy divided by the maximum entropy (Shannon, 1948). In this study, the entropy calculations were performed at the word level (as is visualised by the different columns in Table 2). If all transcriptions of the individual listeners contained exactly the same words, an entropy score of 0 was obtained (as is the case in the column $Word_1$ containing *de* in Table 2). When all entries were different, such as in the last column of Table 2, the relative entropy score was 1. Thus, if all transcriptions are the same, the

entropy score is low which indicates high intelligibility. If the listeners' transcriptions do not agree, the entropy score is higher, and if there is no agreement at all between the transcribers, entropy equals 1.

(1) $$Entropy = \frac{-\sum_{i=0}^{n}(p_i \log_2(p_i))}{\log_2(N)}$$

where p_i = the probability of each word's occurrence; n = the total number of occurrences and N = the number of participants

Three aspects influence the word entropy score: the degree of variance between the transcriptions (i.e., the number of different words in a column), the number of blank spaces and the number of Xs. If listeners identified different words in a particular position in the utterance, this leads to a higher entropy score. When the number of alternative transcriptions increases, the entropy increases by definition, due to the nature of the computation of entropy according to the equation in (1). Blank spaces (if they occurred more than once) on the other hand indicate that the transcriptions agree on the absence of a word in a particular position. Thus, those listeners agreed on the absence of a particular word, while some other listener(s) identified a particular word at that position in the utterance.

Xs are a different matter. X indicates that the listener is not able to identify a particular existing Dutch word. If several X's were aligned, this meant that the listeners agreed that at that position in the utterance, an unidentifiable word occurred. So the agreement between the listeners pertained to the unidentifiability, and hence, unintelligibility of the word uttered by the child. But the agreement did not relate to the identity of the word uttered by the child. For instance, the first

column of Table 2 contains five times the same word, hence the transcribers identified the same word and the entropy equals zero. If that column would have contained five times the symbol X, the same entropy would result, indicating the same degree of agreement between the transcribers. Nevertheless, in the first case the agreement pertains to a specific lexical item that was transcribed identically by all transcribers, while in the second case, the agreement pertains only to the fact that the transcribers were not able to identify a particular word. In order not the inflate the agreement between listeners in the case of Xs, all Xs aligned in a particular column in a datatable such as Table 2, were recoded as unique entries (e.g., X_1, X_2, …).

After calculating the relative entropy score for each column, these scores were averaged per utterance resulting in the final utterance entropy score. This numerical utterance entropy score was used as the dependent variable in the statistical analyses.

Transcriber #	$Word_1$	$Word_2$	$Word_3$	$Word_4$	$Word_5$
Transcription participant 1	de	jongen	ziet	een	kikker
	the	*boy*	*sees*	*a*	*frog*
Transcription participant 2	de	jongen	ziet	de	X
	the	*boy*	*sees*	*the*	*X*
Transcription participant 3	de	jongen	zag	-----	kokkin
	the	*boy*	*saw*	-----	*cook*
Transcription participant 4	de	jongen	zag	geen	kikkers
	the	*boy*	*saw*	*no*	*frogs*
Transcription participant 5	de	hond	zoekt	een	kind

	the	*dog*	*searches*	*a*	*child*
Entropy score	**0**	**0.3109**	**0.6555**	**0.8277**	**1**
Mean Entropy score = 0.5588)					

Table 2: Example of five aligned transcriptions, the corresponding entropy scores per column and the mean entropy score for the utterance

Rating task

In the rating task the listeners indicated the relative intelligibility of each utterance on a scale from "completely unintelligible" to "fully intelligible". The position on the rating scale was transformed automatically by the Qualtrics software into a natural number between 0 and 100. These scores were standardized (converted into a z-score) in order to take into account the idiosyncratic differences between individual participants' rating behaviors. The resulting z-scores were entered into the statistical analyses as dependent variables.

Statistical analyses

The statistical analyses were performed in JMP® Pro 15.2. More specifically, multilevel models (MLM) were applied. This type of statistical approach is especially suited for hierarchically structured data. For this study, hierarchy meant that the utterances originated from individual children, which were at their turn nested in a hearing status (NH or CI). Building the best fitting MLM model is an iterative process in which random effects and fixed effects are successively entered into a model. After adding an effect, a likelihood ratio test was used to assess whether the addition of that factor led to a significantly better fit of the model. If that was the case, that

effect was left in the model, otherwise it was removed. Only the best fitting model is discussed in the results section and included in the tables.

In this study, the main factor of interest was Hearing status (with values CI and NH) and also a (linear or quadratic) effect of chronological age was controlled for. The quadratic effect was included in order to test whether the children's intelligibility scores reached a plateau. If the quadratic effect did not lead to a better fit of the model, it was discarded and not reported in the tables with statistical results. Since the utterances were divided into five series, the factor Series was consistently entered as the first fixed effect. However, adding this factor never lead to a better fitting model. Therefore, it was left out of the analyses. The same holds for the factor Gender. In the analyses in which the CI group was considered separately, fixed effects pertaining only to that group were tested. These factors are the ones referred to in Table 1, viz. Age at Implantation, Hearing Age, Etiology (the cause of the hearing impairment: Genetic, CMV infection, Unknown), Bilateral versus Unilateral CI, Aided and Unaided PTA. If adding these fixed effects did not result in a significantly better model fit, their estimates will not be reported in the in the tables in the results section. In all analyses, results were considered significant when $p < 0.05$.

In the second part of the results section, individual entropy scores were estimated from the null model, i.e., a model with only the random effect of the individual children without any predicting variables. In this way, the deviation of each child from the intercept was computed and represented in a boxplot.

Results

Intelligibility scores for children with CI and NH: entropy

The main question of this study is whether the intelligibility of normally hearing (NH) children's spontaneous speech differs from that of children with a cochlear implant (CI). In this analysis, the intelligibility is represented by entropy scores.

In the first instance, the observed values are inspected. The distribution of all the observed entropy scores has a mean of 0.18 (SD=0.19). The mean entropy score of one child (0.72) falls outside the range determined by the interquartile rule. Hence this child, referred to as CI2 in Table 2, can be considered as an outlier and will not be further considered in the statistical analyses. Inspection of the observed values reveals that the entropy of the transcriptions of the children with NH is considerably lower than the entropy of the children with CI. For the NH children: mean = 0.13, SD = 0.12, 95% CI: 0.11 – 0.15, median = 0.09, IQR = 0.18. For the CI children: mean = 0.21, SD = 0.19, 95% CI: 0.18-0.24, median = 0.15, IQR = 0.24. This suggests that the transcriptions of the NH children are considerably more uniform than those of the CI children. Moreover, the variation between the CI children is much larger than that between the children with NH, judging from the standard deviation and the interquartile range of the distributions of their entropy scores.

The best fitting model for the data is reported in Table 3 and consists of the fixed effects Hearing status (NH versus CI) and Chronological age (centred at 85 months). Moreover, the individual children as a random effect improves the model significantly ($p = 0.013$). The model shows that

the entropy scores of children with NH and CI differ significantly ($p = 0.006$). More specifically, the entropy score of children with CI is significantly higher than that of children with NH, meaning that the transcriptions of children with CI's samples show less agreement between the listeners. The estimated entropy score for NH children at intercept is 0.18 and for children with CI 0.22. Considering that the entropy scores can range from 0 to 1, it appears that NH as well as CI children show relatively low entropy scores suggesting relatively high intelligibility of both groups of children.

Furthermore, the best fitting model shows a significant linear effect of chronological age ($p = 0.001$). This means that an increase in chronological age leads to a significant decrease of the entropy score (as visualised in Figure 1), suggesting that older children reach lower entropy scores (and thus higher intelligibility) than the younger children in the sample. Thus, speech intelligibility improvements seem to continue into advanced childhood (primary school age). The quadratic effect of chronological age was not significant and did not lead to a better fitting model, implying that no floor effect was estimated. An interaction between the factors Hearing status and Chronological age did not lead to a better fitting model and, hence, is not reported in Table 3. The lack of an interaction effect between hearing status and chronological age suggests that the change in entropy score relative to chronological age is comparable for children with NH and CI ($p > 0.05$).

	Estimate	**Std. Error**	**t-ratio**	**p**
Intercept	0.176	0.014	12.616	< 0.0001
Hearing status [CI]	0.041	0.014	2.975	0.006

Chronological age	–0.006	0.002	–3.565	0.001

Table 3: Fixed effects on entropy scores for NH and CI children (fixed effects = Hearing status (CI and NH (= reference category)) and Chronological age; random effect = individual children)

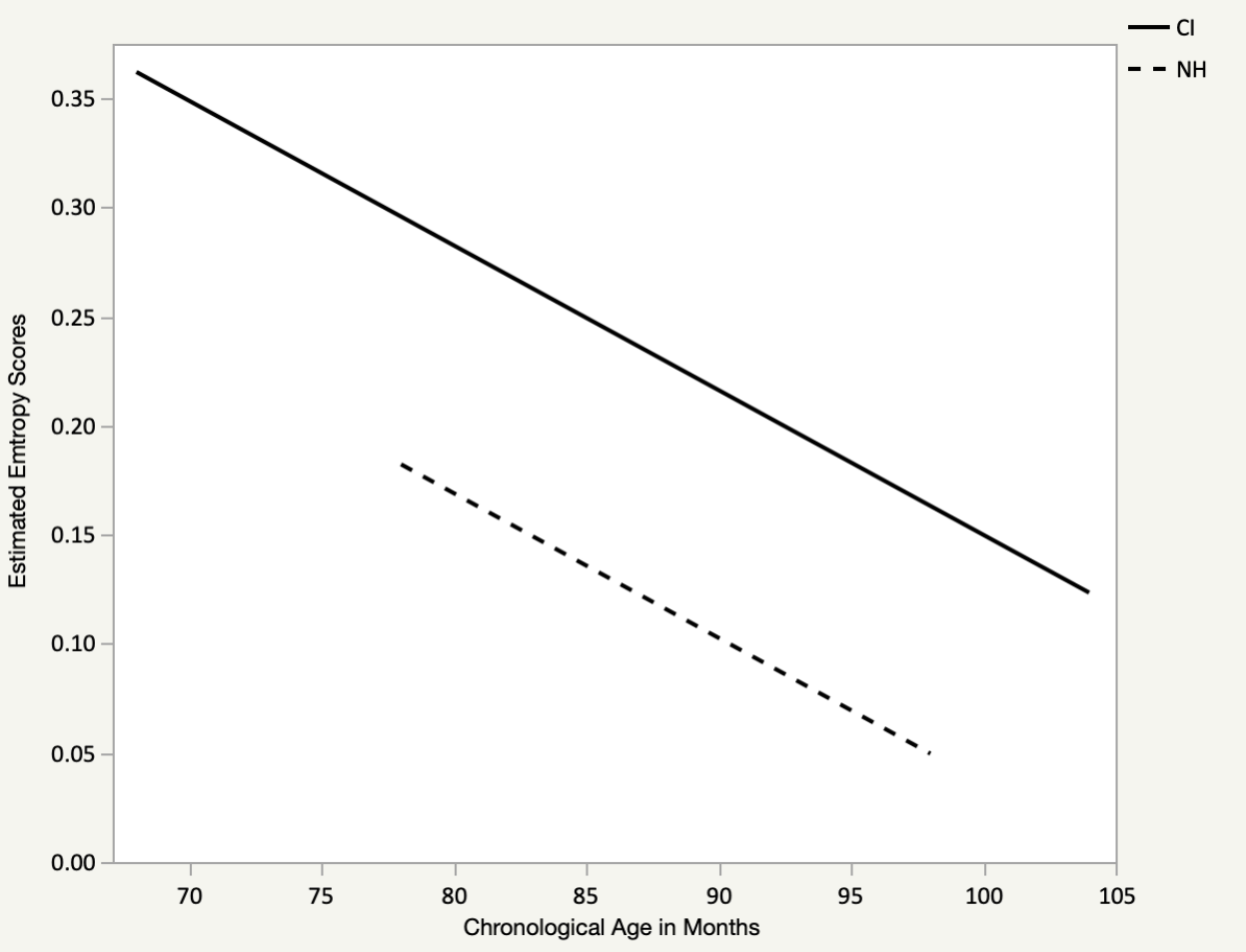

Figure 1: Estimated entropy scores for NH and CI children as a function of their chronological age

In order to assess the development of children with CI and to investigate the effect of the demographic variables specific for this group, a separate model was constructed. As in the

previous analysis, the best fitting model contained the variable Chronological Age. Adding other variables to the model, including Hearing Age, Age at Implantation, Gender, Etiology, (Un)aided PTA, or Bilateral versus Monolateral implants, did not ameliorate the model fit, and hence, did not explain a significant portion of the variance.

Intelligibility scores for children with CI and NH: Rating scales

An analysis of the distribution of the scores on the rating scale, child CI2 shows a discrepant score, viz. a mean score of 20.5 (SD = 10.5) on a scale from 0 to 100, and for the entire group of children the mean score is 62.1 (SD = 20.5). This means that this child can be considered as an outlier and was further discarded in the statistical analyses. The observed (not standardized) values of the ratings of the intelligibility of the children with NH are considerably higher than those of the children with CI. For the children with NH: mean = 69.45 (SD = 17.03), 95% CI 66.79-72.11, median = 72.45, IQR = 27.88, and for the children with CI: mean = 57.06 (SD = 19.58), 95% CI = 53.90-60.22, median = 58.43, IQR = 31.65. As was reported for the entropy scores, the ratings for the two groups of children differs considerably.

The best fitting model for the data is similar to the one for the entropy scores, viz. Hearing status and Chronological age are the significant predictors, as shown in Table 4. As was the case for the entropy scores, adding a quadratic effect of Chronological age did not improve the model, and neither did the interaction of the predictors Hearing status and Chronological age.

	Estimate	**Std. Error**	**t-ratio**	**p**

Intercept	0.005	0.064	0.073	0.942
Hearing status [CI]	-0.219	0.034	-3.441	0.002
Chronological age	0.031	0.008	3.844	0.001

Table 4: Fixed effects on z-score converted scores on the rating scale (fixed effects = Hearing status (CI and NH (= reference category)) and Chronological age; random effect = individual children)

Considering only the ratings of the children with CI reveals a similar pattern of the results for the entropy scores: only chronological age is a significant predictor. Adding the other demographic variables, viz. Gender, Etiology, bilateral versus monolateral CI, Hearing Age and Age at Implantation, did not lead to a significantly better fit.

Individual differences

The previous section indicated that adding the individual children as a random effect significantly improved the model estimating the entropy scores. In order to look into the variability of the individual children, an estimated entropy score is calculated for each child in the sample which is visualised in Figure 2. These scores represent the BLUPs, the best linear unbiased predictions (Henderson, 1975; Liu et al., 2008).

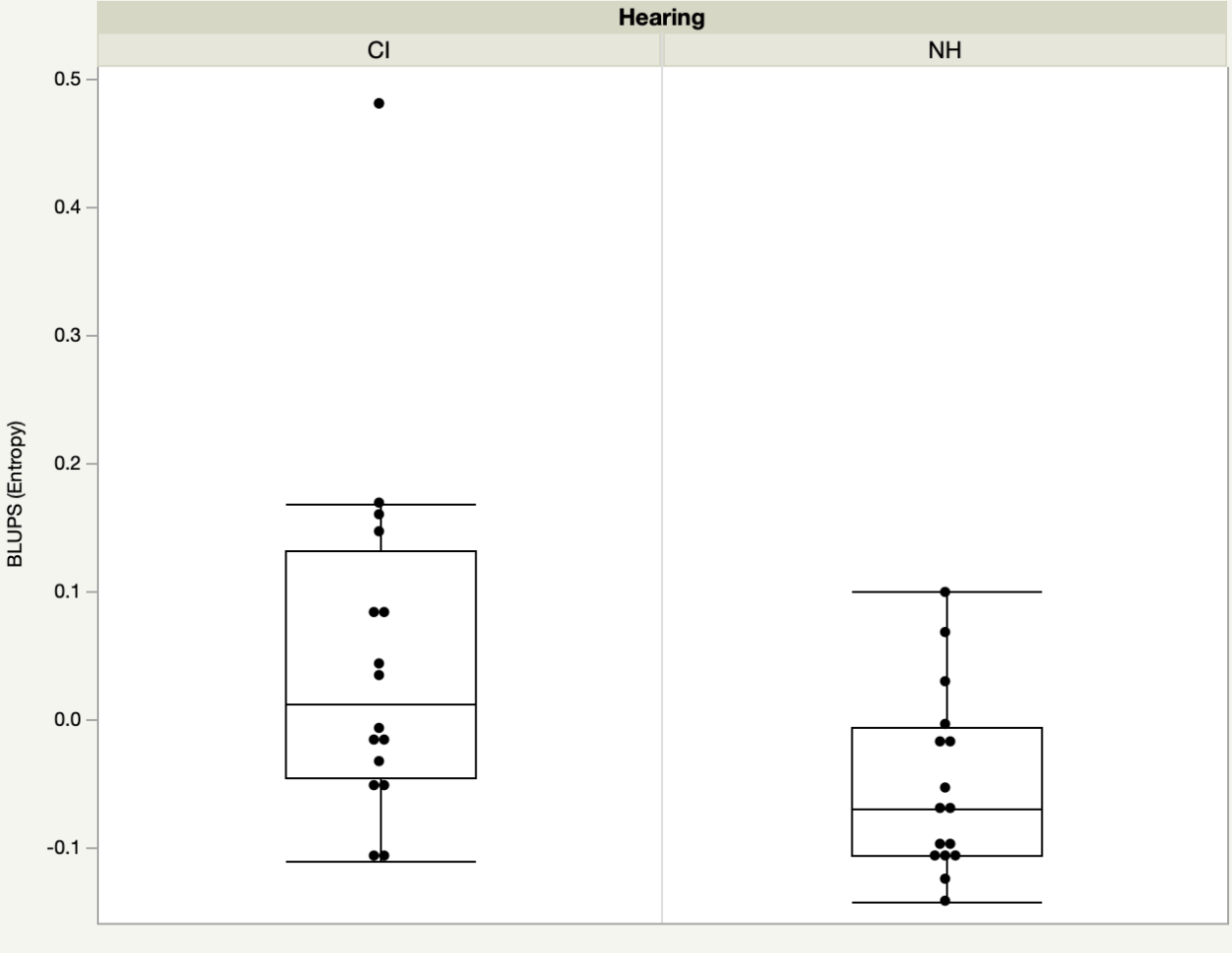

Figure 2: Estimated entropy scores (BLUPS) for the two groups of children based on individual estimated scores (each dot represents the estimated intelligibility score of an individual child)

For the group of NH children, the median estimated entropy score is 0.11 (range: 0.03-0.30). For children with CI, the first striking observation is the outlier in the distribution of entropy values. One child has an average estimated entropy of 0.72, which is almost double of the highest score of the other children. As mentioned in the previous section, this outlier was not included in the statistical modelling. Leaving this outlier out of the analysis, the median entropy score for the children with CI is 0.17 (range: 0.06-0.37). The individual scores of the children with CI show a

larger amount of intra-group variability. However, eight children with CI score below the third quartile of the NH scores, and 12 children obtain scores below the fourth quartile of the NH children. Four children (i.e., CI1, CI2, CI10 and CI12) have intelligibility scores outside the distribution of the NH children, i.e., scores above the 4th quartile.

Correlation of entropy and scale scores

In the present study, 100 participants transcribed utterances of the two groups of children and 150 participants rated the intelligibility of the same utterances. The rating was holistic in the sense that the participants listened to the stimuli and then positioned a slider between the extremes "fully unintelligible" and "fully intelligible". The resulting position of the slider was then projected on a scale between 0 and 100 by the Qualtrics software. It was hypothesized if both tasks tapped onto the same reality, viz. the intelligibility of the children's spoken utterances, a high correlation should surface in a correlational analysis. More specifically, since a high entropy score indicates an elevated level of divergence of the transcriptions, and hence, low intelligibility, a negative correlation was expected with the score on the rating scale. Indeed, low intelligibility corresponds with a low value on the rating scale, and conversely, high intelligibility with a high value on that scale.

A correlational analysis confirms this expectation: a pairwise correlational analysis of the entropy scores resulting from the transcription task and the z-score converted ratings on the scale yields a high negative correlation (Pearson production-moment correlation = -0.906, $p < 0.0001$). This shows a significant linear relationship between the two variables. Further analysis reveals

that the best relationship is a quadratic one, which is shown in the scatterplot in Figure 3 in which for the sake of familiarity the raw scores are represented on the X-axis.

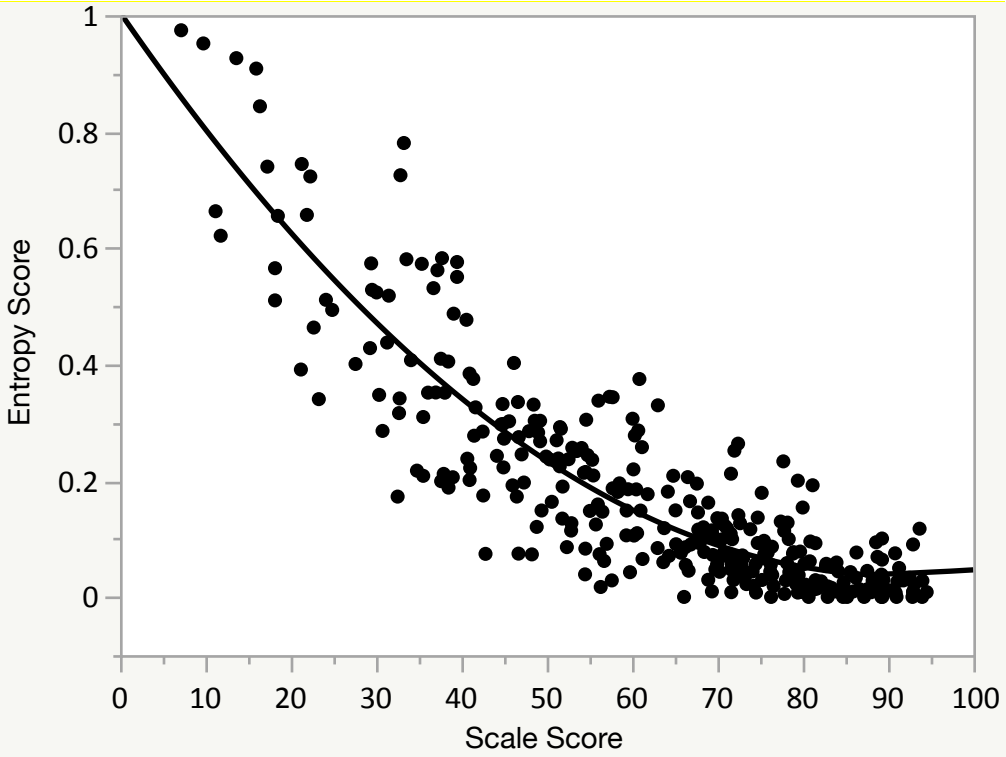

Figure 3: Scatterplot of the entropy scores relative to the scale scores.

The quadratic relationship between the Entropy score and the Scale Score (SS) is expressed in equation (2):

(2) $\text{Entropy} = 0.555 - 0.007*SS + 0.0001*(SS-62.11)^2$

This relationship is highly significant: the R^2 Adjusted equals 0.79, indicating that 79% of the variance in the Entropy score is explained by difference in speech intelligibility expressed by the rating scale. Conversely, 21% of the variance of the entropy is left unexplained by the rating scale.

Discussion

The aim of the present study was to investigate the spontaneous speech intelligibility of seven-year-old Dutch speaking children with CI compared to their chronological age matched NH peers. The children with CI were all early implanted at around 1 year of age. The children's intelligibility was estimated by comparing multiple transcriptions of their speech and computing the entropy of the transcriptions, and by having listeners rate the intelligibility on a perceptual rating scale. The main findings of the study can be summarized as follows. First of all, it was found that the intelligibility of children with CI whose implant was activated around one year of age (the youngest child was implanted at the age of five months), was still lagging behind that of children with NH, even after approximately six years of device use. Secondly, children's intelligibility appears to increase linearly with age. That is, older children were more intelligible for the listeners than younger ones. This effect was apparent in the group of children with CI as well as in the group with NH, indicating that between approximately six and eight years of life children's intelligibility still increases significantly. Moreover, the linear effect of age and the lack of a significant quadratic effect of age suggests that their intelligibility has not reached a ceiling level yet.

The third finding concerns to the method for measuring the intelligibility of spontaneous speech. Children's intelligibility has predominantly been studied using highly controlled speech, as in imitation studies. Spontaneous speech was deemed out of reach because an objective basis for judging their productions as correct or not, was lacking. By using multiple transcriptions of children's spontaneous speech samples and by computing the entropy of those transcriptions, a

method was implemented for assessing intelligibility without assuming a “correct” transcription. A fourth noteworthy finding is that this "analytic" approach of speech intelligibility correlated in a significant way with assessments using “holistic” judgements on a rating scale, thus providing an empirical validation of the approach using entropy.

Intelligibility of children with CI in comparison to NH children

The present study shows that the entropy score and the perceptual ratings were significantly higher for NH children than for children with CI. In other words, NH children’s intelligibility appears to be higher than that of children with CI. For both groups, there was an effect of chronological age, which means that intelligibility increases as children grow older. The effect of chronological age established in the present study corroborates the findings of other studies in which age was found to correlate with language outcomes including intelligibility (Boons et al., 2013; Chin et al., 2003; Flipsen & Colvard, 2006). Remarkably, the intelligibility of neither of the two groups reached a plateau, as can be inferred from the lack of a significant quadratic effect of age. However, it is yet unknown if and when children reach maximal intelligibility. Hustad et al. (2020) estimated the intelligibility of four-year-olds at around 78% in an imitation task. In the present study the average entropy score at approximately eight years of age is predicted to be 0.04 on a scale from 0 to 1, which almost tops a perfect score. But for children with CI the estimated entropy score is still considerably higher, viz. 0.15. In this respect Miller (2013: 606) already noted “that even ‘healthy’ speakers do not achieve 100% intelligibility”. In the present study also NH children of approximately seven years of age did not score a 100% intelligibility score. The question then is: what is maximal intelligibility? What is the level of intelligibility that "healthy speakers", in Miller's words, eventually are able to reach and at which

age do they reach that point? Since the sample in this study only contained a single recording per child and were not selected from a longitudinal follow up of the same children, the effect of chronological age can only be interpreted as: older children are more intelligible than the younger ones. Hence, at this point a longitudinal follow-up is called for in order to confirm that children's speech intelligibility still continues to improve up to and after age seven.

The findings of the present study corroborate those reported in the literature concerning the effect of age on children's intelligibility: irrespective of their hearing, children's intelligibility increases as they grow older, but at a particular age CI children's intelligibility lags behind that of NH children (Chin & Kuhns, 2014; Freeman et al., 2017). Moreover, the variability among children with CI is much larger than that among NH children. This can easily be inferred from the results of the present study (see Figure 2). However, some caution is also required in interpreting the results. On the one hand, of the children with CI participating in the present study, twelve score within the range of the NH children, and the score of only four of them is outside that range, including an obvious outlier. This result seems to corroborate the findings of an increasing number of studies which show that early implanted children are catching up with their NH peers after a few years of device experience (Boons et al., 2013; Geers & Nicholas, 2013; Habib et al., 2010; Nicholas & Geers, 2007; Wie, 2010). On the other hand, the children with CI participating in this study are not an unbiased sample of congenitally hearing-impaired children with a CI. The present sample consists of children with an early detected hearing impairment, who were implanted at an early age, with no additional comorbidities, with parents belonging to the mid-to-high SES, etc. These are all characteristics which have been shown to be favourable circumstances for speech and language development.

The relative homogeneity of the sample of children with CI probably explains some unexpected findings of the present study. First of all, a factor which has been shown time and again to influence the outcome of children's speech and language development is the age at implantation (i.a. Boons et al., 2012; Niparko et al., 2010). The analyses presented here show that the age at implantation is not a significant predictor of the children's intelligibility at age seven, contrary to the findings presented by i.a. Habib et al. (2010). This seems to suggest that it is not the age at implantation, but the children's experience with their implant, i.e., their hearing age, which determines more strongly their intelligibility. But also that did not turn out to be the case: length of device use was not a significant predictor in the analyses presented here. It was the children's chronological age which determined the entropy of the transcriptions and the scores on the rating scale most significantly. At present we can only speculate about the relative effect of these factors. The fact that chronological age was found to be a significant predictor of intelligibility and not age at implantation or hearing age, may be interpreted as indicating that given the small range of the age at implantation of the children studied here and given the small range of their hearing age, the variability was too small in order to exert a significant effect. But alternatively, it may be the case that after a certain amount of time, the effect of the age at implantation is simply not significant anymore, and other factors take over that role, as advocated by Szagun & Stumper (2012).

Using entropy to measure the intelligibility of spontaneous speech

Intelligibility has been mainly measured using (highly) controlled speech in studies of children's speech and especially in clinical studies. Participants (patients) were typically instructed to read a list of words or sentences. Or participants were instructed to repeat or imitate words or sentences read to them. In such a procedure, the researcher typically judges each word or sentence as either correct or incorrect and uses some summary statistics to quantify the level of intelligibility of the participant's speech as, for instance, the percentage of words repeated correctly. The main advantage of such a procedure is that the target is clearly determined in advance: the list of words or sentences to be read or repeated is the target and the participant's rendition can be compared with that target. However, in spontaneous, conversational speech, the target of the speaker is in principle unknown, unless the investigator addresses the participant's introspection, which is obviously difficult, if not impossible, in the case of young children. Thus, there is no prespecified target with which a child's spontaneous production can be compared with. This makes a transcription task hazardous: how to rate a transcribed word as (in)correct, if the target is unknown?

In the literature the lack of a target has been addressed by using rating scales on which the child's intelligibility is situated relative to two extremes, such as a Likert scale with the extremes "fully unintelligible" and "fully intelligible", or a scale on which the various grades are labelled as is the case for the SIR (Cox & McDaniel, 1989). In all of these cases, intelligibility is graded in a "holistic" way: irrespective of the (unknown) target, what the child says is evaluated relative to an implicit scale of intelligibility. The alternative approach proposed here, takes as its starting point the child's speech production and several transcribers produce a transcription. The assumption is that transcribers will agree on what the child says if the utterance is intelligible and

will disagree more relative to declining intelligibility. The methodology proposed in this study is to compute entropy as a quantitative expression of the degree of consensus or the degree of chaos among multiple transcriptions. Entropy takes into account the degree of agreement between transcriptions, but also the degree of disagreement between transcriptions (how many different items occur in the transcriptions?) as well as the distribution of those agreements and disagreements. As such, entropy is not only a suitable measure for transcriptions of spontaneous speech, but also for transcriptions of read or imitated speech, especially to shed light on the *degree* of (un)intelligibility of speech samples. For instance, when imitated speech is transcribed, the intelligibility is usually expressed as the percentage of correctly identified words (relative to the total number of words). The decision is binary: a word is either correctly or incorrectly identified. If the target is, e.g., "frog", all instances of "frog" are labelled as "correct", while alternative transcriptions are labelled as "incorrect". But incorrect instances are not further taken into account, which obscures to a considerable extend the degree of intelligibility since the score remains the same whether or not an incorrectly transcribed word is rendered in exactly the same way by the transcribers or in various different ways. For instance, suppose that a particular word is transcribed correctly in 50% of the cases. What does the remaining 50% of the transcriptions consist of? Possibly the remaining 50% of the transcriptions contains exactly the same word so that there are only two variants in the transcriptions (e.g., the correct transcription "frog" and an incorrect one, such as "frogs"). But it is also conceivable that all the incorrect transcriptions are different words (e.g., "frogs", "fox", "fog", etc.). In both cases, the percentage correct is 50%, but the entropy score will be markedly different in both cases. In the case in which there are only two different forms in the transcriptions, the entropy score is still fairly small. However, in the case of the second scenario, the entropy score is greatly affected by the number of different or

even unique transcribed words (see for instance the rightmost column in Table 2), and the entropy score will be fairly high. Hence, these different scenarios are reflected in differences of the entropy score. Thus, the use of entropy to measure provides a fine grained metric of speech intelligibility that goes far beyond what traditional methods have provided.

Interestingly a high correlation was established between the “objective” measurement of intelligibility based on the entropy of transcriptions and the "subjective" holistic measures provided by rating on an unlabelled scale. In the current study, a highly significant correlation of r=-0.85 was computed. This corelation is higher than the one reported by Habib et al. (2010), viz. r=0.79, but lower than the one reported by Peng et al. (2004), viz. r=0.91. This implies that both measures estimate the same reality but to a different extend. They do not measure exactly the same variable. In the transcription task transcribers identify and transcribe words in the child’s speech. The metric measures the degree of agreement between different transcribers' identifications. In the rating scale approach, identification of the linguistic items probably plays an important role but that is not necessarily the case. More and different information can be taken into account in addition to the identification of words, such as the child’s quality of voice, articulatory features such as accuracy, regional accent, and the like. The fact that such ratings use implicit criteria of intelligibility make them less open for a more accurate and explicit assessment.

Perspectives for future research

The present study was restricted to computing the entropy of transcriptions and relating those measurements to particular explanatory variables, such as the children's hearing status and their chronological age. However, the question turns up which specific linguistic or acoustic variables explain particular entropy values. For example, does the level of entropy of the transcriptions produced by listeners increase or decrease given particular phonetic or phonological variables, or other linguistic variables such as certain word types or utterance length? In other words, what are the linguistic determinants of entropy values?

A preliminary qualitative investigation of our data revealed discrepancies between transcribers at different levels. For instance, at the segmental level differences of voicing of the same segment in transcriptions of the same word were found. For example, boom [**b**oːm] *'tree' versus* pomp [**p**ɔmp] *'pump'*. Or differences in the place of articulation between the transcriptions of listeners, e.g., hen [ɦɛ**n**] *'them' versus* hem [ɦɛ**m**] *'him'*, or gaat [**ɣ**aːt] *'goes'* versus had [**ɦ**ɑt] *'has'*). And listeners identified different vowels (including diphthongs) at the same position, e.g., bijen [b**ɛj**ən] *'bees'* versus buien [b**œy**ən] *'shower',* as well as consonants, e.g., was [**w**ɑs] *'was'* versus valt [**v**ɑlt] *'fell'*). It should be noted that for these kinds of discrepancies a phonemic transcription is obviously more appropriate than an orthographic one, as was used in the present study. Morphological differences were also apparent. In our sample word endings were often deviant, as in kikker *'frog' versus* kikker**s** *'frogs'*, schoen *'shoe'* versus schoen**en** *'shoes',* sta *'stand'* versus staa**t** *'stands'*.

These differences between transcriptions may be used in a more refined calculation of entropy. In the present study, each deviance of the transcriptions was equally weighed. In other words,

each difference equally increased the entropy score. However, some deviances in the transcriptions are fairly small (e.g., kikker *'frog'* vs. kikker**s** *'frogs'*), whereas others can really be considered as mismatches (e.g., jongen *'boy'* vs. hond *'dog'*). Further research is needed for finding fruitful ways to refine the measure by taking into account the (linguistic) distance between different transcriptions. For example, the orthographic transcriptions of the listeners could be converted to and aligned on a phonemic level and calculating entropy could take into account the phonological distance of the different alignments (Faes et al., 2016).

Conclusion

This study investigated the spontaneous speech intelligibility of seven-year-old normally hearing (NH) children and children with a cochlear implant (CI). Intelligibility scores were calculated using transcription entropy, i.e., a measure of the degree of chaos among listeners' transcriptions. In addition, intelligibility was holistically judged on a rating scale. A first conclusion is that the intelligibility of the early implanted children with CI was significantly lower as that of their normally hearing peers, implying that they have not caught up with their NH peers yet. Despite the group differences between children with NH and CI, a remarkable result of this study is that there is a high degree of overlap between both groups when considering the children as individuals rather than a group: a majority of the children with CI reach intelligibility scores within the range defined by the NH children. A second conclusion is that speech intelligibility still seems to develop further still seems to continue over time. In both groups of children, older children reach higher levels of intelligibility than the younger ones.

References

AlSanosi, A., & Hassan, S. M. (2014). The effect of age at cochlear implantation outcomes in Saudi children. *International Journal of Pediatric Otorhinolaryngology,* 78(2), 272-276. 10.1016/j.ijporl.2013.11.021

Barnard, J., Fisher, L., Johnson, K., Eisenberg, L., Wang, N.-Y., Quittner, A., Carson, C., Niparko, J. (2015). A prospective longitudinal study of U.S. children unable to achieve open-set speech recognition 5 years after cochlear implantation. *Otology & Neurotology, 36*(6), 985-992. doi:10.1097/MAO.0000000000000723

Baudonck, N., Dhooge, I., & Van Lierde, K. (2010). Intelligibility of hearing impaired children as judged by their parents: A comparison between children using cochlear implants and children using hearing aids. *International Journal of Pediatric Otorhinolaryngology, 74*, 1310-1315. doi:10.1016/j.ijporl.2010.08.011

Baudonck, N. L. H., Buekers, R., Gillebert, S., & Van Lierde, K. M. (2009). Speech intelligibility of Flemish children as judged by their parents. *Folia Phoniatrica et Logopaedica,* 61(5), 288-295. doi: 10.1159/000235994

Bavin, E., Sarant, J., Leigh, G., Prendergast, L., Busby, P., & Peterson, C. (2018). Children with cochlear implants in infancy: predictors of early vocabulary. *International Journal of Communication Disorders, 53*(4), 788-798. doi: 10.1111/1460-6984.12383

Boons, T., Brokx, J. P., Dhooge, I., Frijns, J. H., Peeraer, L., Vermeulen, A., Wouters, J., & van Wieringen, A. (2012). Predictors of spoken language development following pediatric cochlear implantation. *Ear & Hearing,* 33(5), 617-639. doi: 10.1097/AUD.0b013e3182503e47

Boons, T., De Raeve, L., Langereis, M., Peeraer, L., Wouters, J., & van Wieringen, A. (2013). Expressive vocabulary, morphology, syntax and narrative skills in profoundly deaf children after early cochlear implantation. *Research in Developmental Disabilities,* 34(6), 2008-2022. doi: 10.1016/j.ridd.2013.03.003

Bowen, C. (2011). Table1: Intelligibility. Retrieved from http://www.speech-language-therapy.com/

Bruijnzeel, H., Ziylan, F., Stegeman, I., Topsakal, V., & Grolman, W. (2016). A systematic review to define the speech and language benefit of early (<12 months) pediatric cochlear implantation. *Audiology and Neurotology, 21*, 113-126. doi: 10.1159/000443363

Calmels, M.-N., Saliba, I., Wanna, G., Cochard, N., Fillaux, J., Deguine, O., & Fraysse, B. (2004). Speech perception and speech intelligibility in children after cochlear implantation. *International Journal of Pediatric Otorhinolaryngology,* 68(3), 347-351. doi: 10.1016/j.ijporl.2003.11.006

Castellanos, I., Kronenberger, W. G., Beer, J., Henning, S. C., Colson, B. G., & Pisoni, D. B. (2014). Preschool speech intelligibility and vocabulary skills predict long-term speech and language outcomes following cochlear implantation in early childhood. *Cochlear Implants International,* 15(4), 200-210. doi: 10.1179/1754762813y.0000000043

Chin, S. B., Bergeson, T. R., & Phan, J. (2012). Speech intelligibility and prosody production in children with cochlear implants. *Journal of Communication Disorders,* 45(5), 355-366. doi: 10.1016/j.jcomdis.2012.05.003

Chin, S. B., & Kuhns, M. J. (2014). Proximate factors associated with speech intelligibility in children with cochlear implants: A preliminary study. *Clinical Linguistics & Phonetics,* 28(7-8), 532-542. doi: 10.3109/02699206.2014.926997

Chin, S. B., & Tsai, P. L. (2001). *Speech intelligibility of children with cochlear implants and children with normal hearing: A preliminary report.* Progress Report, Indiana University, Bloomington, Indiana.

Chin, S. B., Tsai, P. L., & Gao, S. (2003). Connected speech intelligibility of children with cochlear implants and children with normal hearing. *American Journal of Speech-Language Pathology,* 12(4), 440-451. doi: 10.1044/1058-0360(2003/090)

Cox, R. M., & McDaniel, D. M. (1989). Development of the speech intelligibility rating (SIR) test for hearing aid comparisons. *Journal of Speech, Language, and Hearing Research,* 32(2), 347-352. doi: 10.1044/jshr.3202.347

De Raeve, L. (2010). A longitudinal study on auditory perception and speech intelligibility in deaf children implanted younger than 18 months in comparison to those implanted at later ages. *Otology & Neurotology,* 31(8), 1261-1267. doi: 10.1097/MAO.0b013e3181f1cde3

Dettman, S., Dowell, R., Choo, D., Arnott, W., Abrahams, Y., Davis, A., Dornan, D., Leigh, J., Constantinescu, G., Cowan, R.. & Briggs, R. (2016). Long-term communication outcomes for children receiving cochlear implants younger than 12 months: a multicenter study. *Otology & Neurotology, 37*, e82-e95. doi: 10.1097/MAO.0000000000000915.

Drennan, W., & Rubinstein, J. (2008). Music perception in cochlear implant users and its relationship with psychophysical capabilities. *Journal of Rehabilitation Research and Development, 45*, 779-790. doi:10.1682/JRRD.2007.08.0118

Duchesne, L., & Marschark, M. (2019). Effects of age at cochlear implantation on vocabulary and grammar: a review of the evidence. *American Journal of Speech-Language Pathology, 28*, 1673-1691. doi: 10.1044/2019_AJSLP-18-0161

Ertmer, D. J. (2010). Relationships between speech intelligibility and word articulation scores in children with hearing loss. *Journal of Speech, Language, and Hearing Research,* 53(5), 1075-1086. doi: 10.1044/2019_AJSLP-18-0161

Faes, J., Gillis, J., & Gillis, S. (2015). Syntagmatic and paradigmatic development of cochlear implanted children in comparison with normally hearing peers up to age 7. *International Journal of Pediatric Otorhinolaryngology, 79*, 1533-1540. doi: 10.1016/j.ijporl.2015.07.005

Faes, J., Gillis, J., & Gillis, S. (2016). Phonemic accuracy development in children with cochlear implants up to five years of age by using Levenshtein distance. *Journal of Communication Disorders,* 59, 40-58. doi: 10.1016/j.jcomdis.2015.09.004

Fagan, M., Eisenberg, L., & Johnson, K. (2020). Investigating early pre-implant predictors of language and cognitive development in children with cochlear implants. In M. Marschark & H. Knoors (Eds.), *The Oxford handbook of deaf studies in learning and cognition* (pp. 46-59). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780190054045.013.3

Fang, H. Y., Ko, H. C., Wang, N. M., Fang, T. J., Chao, W. C., Tsou, Y. T., & Wu, C. M. (2014). Auditory performance and speech intelligibility of Mandarin-speaking children implanted before age 5. *International Journal of Pediatric Otorhinolaryngology,* 78(5), 799-803. doi: 10.1016/j.ijporl.2014.02.014

Flipsen, P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics,* 20(4), 303-312. doi: 10.1080/02699200400024863

Flipsen, P. (2008). Intelligibility of spontaneous conversational speech produced by children with cochlear implants: A review. *International Journal of Pediatric Otorhinolaryngology,* 72(5), 559-564. doi: 10.1016/j.ijporl.2008.01.026

Flipsen, P., & Colvard, L. G. (2006). Intelligibility of conversational speech produced by children with cochlear implants. *Journal of Communication Disorders,* 39(2), 93-108. doi: 10.1016/j.jcomdis.2005.11.001

Freeman, V., Pisoni, D. B., Kronenberger, W. G., & Castellanos, I. (2017). Speech intelligibility and psychosocial functioning in deaf children and teens with cochlear implants. *Journal of Deaf Studies and Deaf Education,* 22(3), 278-289. doi: 10.1093/deafed/enx001

Frinsel, F., Kingma, A., Swarte, F., & Gooskens, C. (2015). Predicting the asymmetric intelligibility between spoken Danish and Swedish using conditional entropy. *Tijdschrift voor Skandinavistiek,* 34(2), 120-138.

Geers, A. E., & Nicholas, J. G. (2013). Enduring advantages of early cochlear implantation for spoken language development. *Journal of Speech, Language, and Hearing Research,* 56(2), 643-655. doi: 10.1044/1092-4388(2012/11-0347)

Geers, A., Nicholas, J., Tobey, E., & Davidson, L. (2016). Persistent language delay versus late language emergence in children with early cochlear implantation. *Journal of Speech, Language and Hearing Research, 59*, 155-170. doi: 10.1044/2015_JSLHR-H-14-0173

Gillis, S. (2018). Speech and language in congenitally deaf children with a cochlear implant. In A. Bar-On & D. Ravid (Eds.), *Handbook of Communication Disorders: Theoretical, Empirical, and Applied Linguistic Perspectives* (pp. 763-790). Berlin: Mouton De Gruyter. doi: 10.1515/9781614514909-038

Gordon-Brannan, M., & Hodson, B. W. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology,* 9(2), 141-150. doi: 10.1044/1058-0360.0902.141

Grandon, B., Martinez, M.-J., Samson, A., & Vilain, A. (2020). Long-term effects of cochlear implantation on the intelligibility of speech in French-speaking children. *Journal of Child Language, 47*(4), 881-892. doi: 10.1017/S0305000919000837

Habib, M. G., Waltzman, S. B., Tajudeen, B., & Svirsky, M. A. (2010). Speech production intelligibility of early implanted pediatric cochlear implant users. *International Journal of Pediatric Otorhinolaryngology,* 74(8), 855-859. doi: 10.1016/j.ijporl.2010.04.009

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics, 31*(2), 423-447. doi: 10.2307/2529430

Hollingshead, A. (1975). *Four-factor index of social status*. Yale University, New Haven, CT.

Houston, D., Beer, J., Bergeson, T., Chin, S., Pisoni, D., & Miyamoto, R. (2012). The ear is connected to the brain: some new directions in the study of children with cochlear implants at Indiana University. *Journal of the American Academy of Audiology, 23*, 446-463. doi: 10.3766/jaaa.23.6.7

Hustad, K., Mahr, T., Natzke, P., & Rathouz, P. (2020). Development of speech intelligibility between 30 and 47 months in typically developing children: A cross-sectional study of growth. *Journal of Speech, Language, and Hearing Research, 63*, 1675–1687. doi:10.1044/2020_JSLHR-20-00008

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology,* 3(2), 81-95. doi: 10.1044/1058-0360.0302.81

Khwaileh, F. A., & Flipsen, P. (2010). Single word and sentence intelligibility in children with cochlear implants. *Clinical Linguistics & Phonetics,* 24(9), 722-733. doi: 10.3109/02699206.2010.490003

Lagerberg, T. B., Asberg, J., Hartelius, L., & Persson, C. (2014). Assessment of intelligibility using children's spontaneous speech: Methodological aspects. *International Journal of Language & Communication Disorders,* 49(2), 228-239. doi: 10.1111/1460-6984.12067

Lejeune, B., & Demanez, L. (2006). Speech discrimination and intelligibility: Outcome of deaf children fitted with hearing aids or cochlear implants. *B-ENT,* 2(2), 63-68.

Liu, X., Rong, J., & Liu, X. (2008). Best linear unbiased prediction for linear combinations in general mixed linear models. *Journal of Multivariate Analysis, 99*(8), 1503-1517. doi: 10.1016/j.jmva.2008.01.004

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. doi: doi.org/10.21415/3mhn-0z89

Mayer, M. (1969). *Frog, where are you?* New York: Penguin Putnam Inc. doi: 10.1080/01638539909545082

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders,* 48(6), 601-612. doi: 10.1111/1460-6984.12061

Moberg, J., Gooskens, C. S., Nerbonne, J., & Vaillette, N. (2007). Conditional entropy measures intelligibility among related languages. In P. Dirix, I. Schuurman, V. Vandeghinste, & F. Van Eynde (Eds.), *Computational Linguistics in the Netherlands 2006: Selected papers from the 17th CLIN Meeting* (pp. 51-66). Utrecht: LOT.

Montag, J. L., AuBuchon, A. M., Pisoni, D. B., & Kronenberger, W. G. (2014). Speech intelligibility in deaf children after long-term cochlear implant use. *Journal of Speech, Language, and Hearing Research,* 57(6), 2332-2343. doi: 10.1044/2014_JSLHR-H-14-0190

Nicholas, J. G., & Geers, A. E. (2007). Will they catch up? The role of age at cochlear implantation in the spoken language development of children with severe to profound hearing loss. *Journal of Speech, Language, and Hearing Research,* 50(4), 1048-1062. doi: 10.1044/1092-4388(2007/073)

Niparko, J. K., Tobey, E. A., Thal, D. J., Eisenberg, L. S., Wang, N. Y., Quittner, A. L., & Fink, N. E. (2010). Spoken language development in children following cochlear implantation. *Journal of the American Medical Association,* 303(15), 1498-1506. doi: 10.1001/jama.2010.451

Nittrouer, S., Caldwell-Tarr, A., Moberly, A. C., & Lowenstein, J. H. (2014). Perceptual weighting strategies of children with cochlear implants and normal hearing. *Journal of Communication Disorders,* 52, 111-133. doi: 10.1016/j.jcomdis.2014.09.003

O'Donoghue, G. (2013). Cochlear implants — Science, serendipity, and success. *The New England Journal of Medicine,* 369(13), 1190-1193. doi: 10.1056/NEJMp1310111

Osberger, M. J., Robbins, A. M., Todd, S. L., & Riley, A. I. (1994). Speech intelligibility of children with cochlear implants. *The Volta Review,* 96(5), 169-180. doi: 10.1044/jshr.3601.186.

Ozbic, M., & Kogovsek, D. (2010). Voice quality, articulation, nasality, prosody and overall intelligibility in the speech of subjects with hearing impairment. *Croatian Review of Rehabilitation Research,* 46(1), 41-56. doi: 10.1016/j.jcomdis.2012.05.003

Peng, S.-C., Spencer, L. J., & Tomblin, J. B. (2004). Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience. *Journal of Speech, Language, and Hearing Research,* 47(6), 1227-1236. doi: 10.1044/1092-4388(2004/092)

Qualtrics. (2005) (Version December 2018). Provo, Utah, USA. Retrieved from www.qualtrics.com

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal,* 27(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Strömbergsson, S., Holm, K., Edlund, J., Lagerberg, T., & McAllister, A. (2020). Audience response system-based evaluation of intelligibility of children's connected speech – validity, reliability and listener differences. *Journal of Communication Disorders, 87*, 106037. doi:10.1016/j.jcomdis.2020.106037

Subtelny, J. (1977). Assessment of speech with implications for training. In F. Bess (Ed.), *Childhood deafness: Causation, assessment, and management* (pp. 183-194). New York: Grune & Stratton.

Svirsky, M., Chin, S., & Jester, A. (2007). The effects of age at implantation on speech intelligibility in pediatric cochlear implant users: clinical outcomes and sensitive periods. *Audiological Medicine, 5*(4), 293-306. doi: 10.1080/16513860701727847

Szagun, G., & Stumper, B. (2012). Age or experience? The influence of age at implantation and social and linguistic environment on language development in children with cochlear implants. *Journal of Speech, Language, and Hearing Research, 55*, 1640-1654. doi: 10.1044/1092-4388(2012/11-0119)

Toe, D. M., & Paatsch, L. E. (2013). The conversational skills of school-aged children with cochlear implants. *Cochlear Implants International,* 14(2), 67-79. doi: 10.1179/1754762812y.0000000002

Tye-Murray, N., Spencer, L., & Woodworth, G. G. (1995). Acquisition of speech by children who have prolonged cochlear implant experience. *Journal of Speech and Hearing Research,* 38(2), 327-337. doi: 10.1044/jshr.3802.327

van Heuven, V. J. (2008). Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. *International Journal of Humanities and Arts Computing,* 2(1-2), 39-62. doi: 10.3366/e1753854809000305

Van Lierde, K. M., Vinck, B. M., Baudonck, N., De Vel, E., & Dhooge, I. (2005). Comparison of the overall intelligibility, articulation, resonance, and voice characteristics between children using cochlear implants and those using bilateral hearing aids: A pilot study. *International Journal of Audiology,* 44(8), 452-465. doi: 10.1080/14992020500189146

Verhoeven, J., Hide, O., De Maeyer, S., Gillis, S., & Gillis, S. (2016). Hearing impairment and vowel production. A comparison between typically developing, hearing-aided and cochlear implanted Dutch children. *Journal of Communication Disorders, 59*, 24-39. doi: 10.1016/j.jcomdis.2015.10.007

Weiss, C. E. (1982). *Weiss intelligibility test*. Tigard: CC Publications.

Whitehill, T. L., & Ciocca, V. (2000). Perceptual-phonetic predictors of single-word intelligibility: A study of Cantonese dysarthria. *Journal of Speech, Language, and Hearing Research,* 43(6), 1451-1465. doi: 10.1044/jslhr.4306.1451

Wie, O. B. (2010). Language development in children after receiving bilateral cochlear implants between 5 and 18 months. *International Journal of Pediatric Otorhinolaryngology,* 74(11), 1258-1266. doi: 10.1016/j.ijporl.2010.07.026

Wie, O., von Koss Torkildsen, J., Schauber, S., Busch, T., & Litovsky, R. (2020). Long-term language development in children with early simultaneous bilateral cochlear implants. *Ear & Hearing, 41*(5), 1294-1305. doi: 10.1097/AUD.0000000000000851

Yanbay, E., Hickson, L., Scarinci, N., Constantinescu, G., & Dettman, S. J. (2014). Language outcomes for children with cochlear implants enrolled in different communication programs. *Cochlear Implants International,* 15(3), 121-135. doi: 10.1179/1754762813y.0000000062

Young, G. A., & Killen, D. H. (2002). Receptive and expressive language skills of children with five years of experience using a cochlear implant. *Annals of Otology, Rhinology, and Laryngology,* 111(9), 802-810. doi: 10.1177/000348940211100908