Seifert, MI., Morgan, L., Gibbin, S., & Wren, Y. E. (2019). An alternative approach to measuring reliability of transcription in children's speech samples. *Folia Phoniatrica et Logopaedica*. https://doi.org/10.1159/000502324

An alternative approach to measuring reliability of transcription in children's speech samples: Extending the concept of near functional equivalence

Miriam Seifert[1], Lydia Morgan[1], Sarah Gibbin[2], Yvonne Wren[1,3]

1: Bristol Speech and Language Therapy Research Unit, North Bristol NHS Trust, Bristol, UK

2: Speech and Language Therapy Department, North Bristol NHS Trust, Bristol, UK.

3. Bristol Dental School, University of Bristol, Bristol, UK

Short Title: Measuring Reliability of Transcription

Correspondence:

Miriam Seifert

Bristol Speech and Language Therapy Research Unit,

North Bristol NHS Trust,

Southmead Hospital,

Southmead Road,

Westbury-on-Trym,

Bristol, BS10 5NB

Tel: +44 117 4143951, email: Miriam.seifert@cchp.nhs.uk

Keywords: ALSPAC, perceptual analysis, transcription, speech, reliability

ABSTRACT

Aim: To explore a novel and efficient way of calculating transcription reliability of connected speech data using the concept of near functional equivalence. Using this approach, differences between two transcribed phonemes that are nearly phonetically equivalent are disregarded if both reflect two plausible and acceptable pronunciations for the word produced.

Method: The study used transcriptions of connected speech samples from 63 5-year-olds who participated in a large-scale population study. Each recording was phonetically transcribed by two speech and language therapists. Two independent researchers then examined agreement between the two sets of transcripts, marking differences in vowels, consonants, diacritics and identifying segments which represented near functional equivalence.

Results: Overall percentage agreement between the transcripts was 77%. One quarter of the differences between the two transcripts were identified as showing near functional equivalence. When this category was excluded, the transcripts showed 82% reliability.

Conclusion: This study demonstrates the issues to consider when calculating transcription reliability. Other methods are often time intensive and may highlight differences between transcribed units which are audibly very similar and would be negligible in ordinary conversation. Inclusion of the concept of 'near functional equivalence' can result in higher reliability scores for transcription, without loss of rigour.

INTRODUCTION

Phonetic transcription is used routinely in both clinical and research contexts as a means to record an individual's speech output. The visual representation of speech, which is the output of the transcription, enables "the transcriber to determine how effective or proficient the speaker is as a communicator" [1] (p.300). In order to make such judgements, the transcription must be both valid (i.e. be congruent with findings from other types of data obtained from acoustic or physiological measures) and reliable (i.e. remain highly similar when transcribed by two or more different transcribers or at different times by the same transcriber), [2]. Clinically, the accuracy of the transcription is essential to ensure an appropriate intervention plan is made [3]. For research purposes, reliable transcription is required to enable researchers to analyse speech data and facilitate accurate interpretation of a study's findings [4].

Transcription methods

Whilst reliability in phonetic transcription is clearly important, achieving reliability using perceptual data can be difficult. Many factors impact on the final transcript's objectivity [4], including the quality of the data that are being transcribed (for example live versus video versus audio recording) [5], transcriber background training and experience [6] and whether the transcription is broad (recording productions at a phonemic level) or narrow (providing detailed information about phonetic variations).

A further complication is the size and type of the sample being transcribed. Crowdsourcing, a method where large numbers of non-expert listeners are recruited through online platforms, has been utilised in studies investigating perceptual speech outcomes (7). However, listeners are usually only required to rate the speech samples or make simple correct/incorrect decisions about the accuracy of single phonemes or words (8). Achieving reliability across such samples of single word production is likely to be easier than when large samples of connected speech are involved.

While acoustic analysis can help, perceptual analysis has been reported in the literature frequently as the transcription method of choice for large datasets [9, 10]. It is important, therefore, that the method chosen to measure reliability of transcriptions is fully understood and its constraints openly addressed by researchers as well as users of the research [1].

In clinical speech and language therapy, narrow transcription is recommended to capture phonetic differences that often hold significant information about an individual's phonology, that is, their understanding of how sounds are used contrastively in the language they are speaking. Ball et al. [3] describe several clinical examples where narrow transcription helps to guide therapy. One example was the use of the subscript arrow convention to indicate that the child had marked a sliding articulation i.e. [s͢ʃ]. They argued that without the arrow diacritic, i.e. [sʃ], the production would be classed as two fricatives in a cluster, rather than a subtle change in place of articulation within the time scale of one segment. The diacritic provides more accurate information about the child's ability to produce fricatives. Ball and Rahilly [11] also point out that if an English-speaking child devoices /b/ (e.g. /bɪn/) to [p] but produces [p] without aspiration (e.g. [pɪn]), this is much less perceptible, than if the child had retained the aspiration which is present in the usual adult form i.e. [p^{h}ɪn]. In both examples given, the broad transcription underestimates the individual's ability to signal phonological differences.

However, there is consensus in the literature that it is hard to achieve reliability between transcribers when using narrow transcription and agreement will naturally be lower when more symbols are being used [4]. Shriberg and Lof [6] investigated inter-rater (agreement among raters) and intra-rater (consistency of same rater on repeated tests) transcription reliability using consensus transcription. When using broad transcription, they found agreement of 88% for consonants, and 91% for vowels between transcriber teams. In contrast, agreement was reached on only 13% of consonants and 53% of vowels when narrow transcription was used.

Measuring transcription reliability

There are several different methods for measuring transcription reliability. A method frequently cited in the literature [12, 13] is point-to-point percentage agreement, whereby the number of agreements in two transcriptions is divided by the total number of transcribed units. A percentage agreement of 85% or more is typically reported in the literature [6], though Pye, Wilcox and Siren [14] emphasise that this number has "little objective foundation" and should not confirm the integrity of the transcript. This method also fails to account for types of differences, where some phonemes are phonetically closer than others [11], e.g. [d] and [t] differ in voicing only, whereas [g] and [ʧ] differ in voice, place and manner. Additionally, Cucchiarini [4] points out, if the transcribers use a different number of consonants in a word, for example one transcribes a production of the word ‘artist’ as [a:rtɪst] and the other as [a:təst], the percentage of agreement for that word is very low, yet the spoken productions of each of the two transcriptions would sound very similar.

An alternative method requires two or more transcribers to reach agreement through consensus decision making. This approach is less transparent in terms of establishing the significance of the differences in transcripts and in how consensus was reached. Factors such as the transcribers’ status, personality styles and competence can influence decision making in transcription [1]. Moreover, it is possible that consensus won’t be reached or, if the transcriptions have involved a large dataset and/or taken place over an extended period of time, that the original transcribers are no longer available. Bosma Smit et al [15] used a consensus listening approach and a “transcriber selection procedure” in an attempt to reduce error variance between transcribers, when analysing percentage of consonants correct in word lists and conversation samples. Ten experienced speech and language therapists (SLTs) who were blinded to child identity and treatment group transcribed a series of speech samples. Those transcribers whose transcriptions varied by more than 10% using a point-to-point percentage agreement method were not involved in the final study. The study could then confidently report that all five transcribers involved in the final study were within 10% of each other in pair-wise comparisons for the same speech samples.

A third approach takes account of the fact that not all phonetic differences are of equal value. Cucchiarini [4] proposes a system based on Vieregge [16] matrices, which compares two transcribed units by measuring the average difference between each feature. For example, Cucchiarini explains that /t/ and /s/ have commonalities in that they have the same place of articulation and are both voiceless sounds. However, they differ in terms of manner, whereby /t/ is a stop and /s/ is a fricative. In this method, each manner feature receives a score such that /t/ is scored 1 for stop and 0 for fricative while /s/ scores 0 for stop and 1 for fricative. The combined score for difference between these two phonemes therefore is 2. Similarly, the differences between /t/ and /l/ are voice, lateralisation and stop (whereby /l/ is a voiced lateral and /t/ is a voiceless stop, giving the difference between /t/ and /l/ a score of 3 (one point for the difference in voice and one each for the differences in manner features of lateral and stop). Cucchiarini's [4] approach also takes into account diacritics by determining the effect any diacritics would have on productions of the transcribed unit. Ball and Rahilly [11] refer to a similar system when measuring inter-rater reliability, whereby the phonetic features, that is the voice, place and manner of a sound are taken into account and two transcriptions are deemed as a 'complete match', 'match within one phonetic feature' and 'non-match'.

Similar to Cucchiarini [4] above, attempts have been made to classify diacritics by the significance of their differences. Shriberg and Lof [6], who categorised diacritics into 7 different classes including 'nasality', 'stop release', 'tongue position' and 'sound source', propose that diacritic agreement in transcriptions should be categorised as being either exact, having within-class agreement or having any diacritic, disregarding its class. Further, Shriberg et al [17] categorised diacritics into those considered to identify errors and those which represent non-errors. The list of non-errors was derived from consideration of each diacritic against the following criteria (where at least one needed to apply): 1) optional during transcription of casual speech (e.g. unreleased [p̚]), 2) not reliability transcribed and 3) a lay person would not perceive them as an articulation difficulty (e.g. [bæ̃t]). Diacritics in the error list are those that represent non-optional allophones (e.g. nasal emission), are reliably transcribed, and are likely to be considered variations that require intervention (e.g. lateralisation).

Another approach to transcription reliability measurement is proposed by Shriberg and Kent [18]. They also recognise that not all differences in transcriptions of speech samples are of equal value and propose ways of reaching agreement that place more value on the functional aspects of transcription. They refer to ‘functional equivalence’ which they define as “essentially equivalent phonetic transcriptions of a target behaviour that uses alternative symbolization” and provide the example that a lowered /i/ (i.e. [i̞]) and a raised /ɪ/ (i.e. [ɪ̝]) are perceptually very similar but can be represented by two different phonetic symbols by transcribers. They also highlight other examples where two phonemes are ‘nearly functionally equivalent’ which they define as “nearly equivalent phonetic transcriptions of a target behaviour in terms of place and manner features” and provide the example of a [s] and a fricated [t͓]. They propose that any units be compared and categorised as to whether they are ‘identical’, ‘functionally equivalent’ or ‘nearly functionally equivalent’.

Shriberg and Kent’s [18] categorisations are particularly useful when large datasets of connected speech are involved. Transcribing connected speech is important because we mostly do not communicate in single words and connected speech samples provide a more realistic impression of a child’s phonetic and phonological competence. During connected speech, boundaries between sounds, syllables and words are constantly blurred [19] and different components of speech influence each other [20]. There are several common connected speech characteristics, for example, consonants in one word can affect the initial consonant of the next word (assimilation), or the final phoneme in a word can be deleted due to the features of the subsequent word (elision). These features can be difficult to perceive and, as a consequence, difficult to transcribe. This may result in differences between two transcriptions, leading to a low reliability score, when in fact the differences between the two transcripts represent negligible differences in the actual speech produced.

The current paper reports a novel way of analysing transcription reliability data that considers the issue of 'near functional equivalence' and extends the concept through focusing on whether differences in phonetic transcription are likely to be audibly perceptible when spoken. In other words, as well as using the term for two productions which might be considered near equivalent as in the example of [s] and a fricated [t͓] above, the term is applied for those differences between two transcriptions of connected speech where differences reflect two plausible and acceptable pronunciations for a given word. This is based on the tenet that communication takes place in real-life conditions where specific nuances of speech go unnoticed and are often irrelevant to the message that a speaker is trying to convey [21]. It is also anticipated that this approach would increase reliability of transcription without compromising quality.

The study used connected speech samples from 5-year-old children who participated in a large-scale normative population study. The aim of this work was to explore the impact on inter-rater reliability estimates of adopting a 'near functional equivalence' approach to reliability of transcription.

METHOD

Participants

Participants for this study were 5-year-old children who had been recruited to the Avon Longitudinal Study of Parents and Children (ALSPAC). Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled was 14,541 (for these at least one questionnaire had been returned or a “Children in Focus” (CiF) clinic had been attended by 19/07/99). Of these initial pregnancies, there were a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age.

A 10% sample of the ALSPAC cohort, known as the Children in Focus (CiF) group, attended clinics at the University of Bristol at various time intervals between 4 to 61 months of age. The CiF group were chosen at random from the last 6 months of ALSPAC births (1432 families attended at least one clinic). Excluded were those mothers who had moved out of the area or were lost to follow-up, and those taking part in another study of infant development in Avon. The phases of enrolment are described in more detail in the cohort profile paper [22, 23]. Please note that the study website contains details of all the data that are available through a fully searchable data dictionary and variable search tool at http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/.

Data Collection

Speech recordings

1432 children were invited and 988 children attended the CiF clinic at age 61 months (69%). These children were assessed on a wide range of physical, sensory, cognitive and environmental measures. Blood samples were taken and parenting questionnaires completed. Children were also assessed on a range of measures of speech and language. These included a single word naming task adapted from Paden, Novak and Beiter [24] specifically for the clinic; the verbal comprehension subtest of the Reynell Developmental Language Scales-Revised Edition [25]; a test of children's narrative ability (the Renfrew Bus Story Test, [26]); a test of children's ability to identify which two of three words illustrated by line drawings began with the same initial consonants [27]; and a request to repeat two multisyllabic words (butterfly and dinosaur) five times.

Assessors were qualified SLTs. Those elements of the session which required the child to produce speech were orthographically transcribed live during the session and also audio-recorded for later verification. The recordings were made between 1st April 1996 and 31st December 1997. No information on the specification of the equipment used or its set-up was recorded by the study.

The recordings of the Renfrew Bus Story [26] were used as the source for this investigation. Samples of connected speech were preferred to that of single word production as it was considered that this was closer to naturalistic speech used in everyday conversation. The Bus Story test is standardised on children aged 3 to 8 years and was designed as a screening test of verbal expression. It requires children to listen to a story about a naughty bus told with pictures. Children are then asked to retell the story with the picture support. The child's narrative was recorded orthographically and following the assessment, scored for information content and sentence length. Not all children who attended the CiF clinic at 61 months completed all aspects of the speech and language assessment owing to time limitations and cooperation of the child. In total, 162 children refused to cooperate, 779 children completed the Bus Story test and another 47 partially completed it. In total, 826 had connected speech samples. Where necessary, enhancements to increase the audio quality of the recordings were made. However, for 32 cases, the audio quality could not be enhanced sufficiently and transcription was not possible. These recordings were not used in this study. In total therefore 794 samples (80%) were available for transcription and analysis.

Phonetic Transcription

The orthographic transcriptions which had been taken during the assessment were checked against the recordings and errors corrected. All of the recordings were then phonetically transcribed by a qualified SLT. The primary purpose of carrying out these transcriptions was to determine the range of speech production proficiency in this population and to use the scores for this in an analysis to identify risk factors for poor speech outcomes at age 5. Given the size of the dataset, it was not feasible to use narrow transcription throughout due to the time and costs that this would have incurred. As an alternative, transcribers were asked to use broad transcription for most of the speech sample but to use narrow transcription for errors.

As the children in the sample were recruited to a population study, most children had speech which was within the typical range for speech development at age 5. Errors existed as part of typical speech at this age, because the child had a speech impairment or because of idiosyncratic productions in an otherwise typical speaker.

Ten percent of the recordings (77) were selected at random to be phonetically transcribed by another qualified SLT (the first author). Fourteen of these recordings were unavailable at the time of this study. These data were therefore excluded, resulting in 63 transcripts which were used in the final comparison (8% of the sample).

Both transcribers were provided with a list of speech characteristics which are common within the Bristol accent, which is spoken in the geographical area of the study. These included vowels (e.g. [a] for /ɑ:/ as in 'bath'), consonants (e.g. [f] for /θ/) and stylistic variation in all accents (e.g. elision whereby sounds are omitted such as 'expect so' being produced as [spek səʊ]).

Calculating reliability of transcriptions

Two qualified SLTs, independent to those who conducted the transcription, completed the reliability checks. Nearly a third of the reliability checks were conducted by one of the reliability checkers (n=17) and the remaining transcripts (n=46) were checked by the other. In order to ensure reliability between the transcript checks, five of the transcripts were independently assessed by both SLTs.

The two reliability checkers identified differences in the two transcriptions in vowels, consonants and diacritics. They also identified differences which could be classified as 'near functional equivalence'.

For each of the four categories of difference in the transcriptions (vowels, consonants, diacritics and 'near functional equivalence'), the number of differences between the transcript pairs was calculated. The total phonemes for the original transcript were counted using a digital tally calculator. Percentage differences between the samples were then calculated for the vowels, consonants, diacritics and 'near functional equivalence' differences in transcript pairs, as a proportion of the total number of phonemes in the original transcript. For example, if there were seven instances of different vowel symbols used between the two transcripts, and the original transcript contained 243 phonemes, the percentage differences would be calculated as so: 7/243x 100=2.88% differences in vowels across all phonemes in the sample.

Subsequently, the transcript pairs were examined to identify patterns in the differences between each pair. Examples of types of transcription differences that were categorised as 'nearly functionally equivalent' are provided in Appendix A.

RESULTS

In total, 63 transcripts were phonetically transcribed, independently, by the two transcribers. The mean transcript length was 290 phonemes (SD 88, range 84-479).

Of the five pairs of transcripts which were checked by both reliability checkers, differences between the two checkers in their classification of differences were very small. The largest percentage of difference was with vowels (2.3%), 'near functional equivalence' and diacritics had similar differences (1.3% and 1.2% respectively). The smallest difference between the two checkers' classifications was for consonants (0.9%).

Categories of difference in the pairs of transcripts

Table 1 summarises the differences between the pairs of transcripts for each of the categories of difference i.e. vowels, consonants, diacritics and 'near functional equivalence'.

Mean differences for each category are provided together with the range (smallest to largest percentage difference in agreement across the whole sample) and standard deviations. The category with the biggest difference between the two transcribers was consonants, with a mean difference of 9.66%, this was followed by the ‘near functional equivalence’ differences (5.3%) then vowels (4.84%) and finally diacritics (3.43%).

The combined mean total difference between the transcripts, including all categories of difference, was 23%; the overall percentage agreement between the transcripts was therefore 77%. If ‘near functional equivalence’ differences are excluded from analysis, the percentage agreement is 82%. Finally, if diacritics are also excluded, and reliability is considered purely on perceptible consonant and vowel differences, agreement falls within the commonly acceptable level at 85.5%.

Types of difference identified in transcript pairs

Many of the transcription differences that were considered as ‘near functional equivalence’ in connected speech by the reliability checkers, reflected differences related to word boundary features in speech. For example, the phrase ‘the policeman blew’ was transcribed by one transcriber as: [ðə pəlisman blu] and the other as: [d̪ə pəlismam bləʊ]. The difference in transcription of the final consonant of the word ‘policeman’ demonstrates the process of assimilation, whereby the /n/ took on the bilabial place of articulation of the following consonant /b/. It would be very difficult to determine using just perceptual analysis which of these transcriptions should be considered correct.

Other frequent ‘near functional equivalent’ differences in transcription which were observed included the tendency for one transcriber to link vowels with a /j/ (e.g. [taɪjəd] versus [taɪəd] for the word ‘tired’); the use of word final glottal stops versus /t/ (e.g. [went] and [wenʔ] for ‘went’); and the use of syllabic consonants (e.g. [wɪsl̩] and [wɪsəl] for ‘whistle’). Other types of near functional equivalent differences related to: glottal fricatives, clusters, word final n/ŋ, subtle place distinction and word final voicing (see Appendix A). Differences in vowels were often associated with weak vowels such that schwa /ə/ was often alternatively transcribed as /ɒ/, /u/, /ʌ/ and /ɪ/; /ɪ/ itself was alternatively transcribed as /i/; and /ʊ/ as /ʌ/. Differences in vowels were included within this category when they fulfilled criteria for near functional equivalence. Where this wasn’t the case, they were included in the vowel category.

DISCUSSION:

This study explored how calculating ‘near functional equivalence’ could be used as an alternative to reporting simple reliability rates for narrow and broad transcription. Two sets of transcriptions of connected speech from 63 5-year-olds, carried out independently by two SLTs, were compared to determine the level of agreement between each pair of transcripts. When all differences were included in the count, agreement between the transcripts was 77%. However, one quarter of the differences between the two transcripts were identified as showing near functional equivalence and when this category was excluded in the calculations, the transcripts showed 82% reliability.

The present study is based purely on audio recordings, and so details were not able to be checked against visual/video data. Further, no acoustic analysis was conducted to support transcription methods. However, the present study utilises transcribing methods that are frequently used in research and require the least resources.

Some of the features that were noted at the start of this paper as having the potential to affect the objectivity of a transcript may also play a role in the objectivity of comparing transcript reliability. Two individuals carried out the transcript reliability check and a criticism of using the 'near functional equivalence' approach is that it is subjective, requiring individuals to decide what they consider to be a different, yet equivalent sound. Despite this, the present study found high levels of agreement between the reliability checks carried out by the two individuals. There was only 1.3% difference between reliability raters in the 'near functional equivalence' differences group. Since both reliability checkers were qualified SLTs, it is perhaps more likely that these trained professionals will have a shared agreement of what acceptable or equivalent speech sounds are [6]. As such, it is recommended that expert opinion, as utilised in this study, is always used to calculate transcription reliability when using this method.

The existing literature has indicated relatively high levels of agreement between transcribers when using broad transcription, e.g. Shriberg and Lof [6] found 88% agreement for consonants and 91% for vowels. Similar, but slightly higher levels of agreement, were found in the present study with 90% consonant agreement and 95% vowel agreement.

The transcribers in this study were instructed to use narrow transcription for errors only, due to the costs involved in using narrow transcription throughout such a large dataset. Of interest was the variability between the two transcribers in their use of diacritics for the narrowly transcribed segments though, with one transcriber using symbols more frequently than the other. However, it is noteworthy that even when all differences between the two sets of transcripts were included, the overall reliability was relatively high in the present study (77%).

It is interesting to note that the biggest differences between transcripts in the present study was for consonants (10%). Significantly fewer differences were found in the 'near functional equivalence' group (5.3%), vowels (4.8%) and diacritics (3.4%). That the number of differences considered 'near functional equivalence' across all categories, was similar to the number of vowel differences, demonstrates that the number that was classified into this group was relatively small. However, nearly a quarter of the differences were classed in the 'near functional equivalence' group, and this difference is important in terms of the overall acceptable level of reliability between transcribers. When 'near functional equivalence' sounds and diacritics were excluded from the calculation of reliability, agreement between transcribers was 85.5%, which is within the commonly considered acceptable range for transcription agreement. However, if the 'near functional equivalence' differences are included, the agreement falls to below 80%. We would argue that the former approach, i.e. only counting differences in transcription of consonants and vowels differences which would not be classified as 'near functional equivalence', is the most useful way to examine reliability.

In the introduction, it was noted that other systems of comparing transcription reliability which are similar to a 'near functional equivalence' approach, take account of the fact that not all phonetic differences are of equal value. Cucchiarini [4] and Ball and Rahilly [11], both describe systems where sounds are classified by the extent that they match. These approaches provide us with the most detail about the extent of differences between transcripts and are therefore arguably the most robust. However, such approaches are time consuming and, though they provide detailed information about similarities and differences between sounds, they do not indicate whether the differences have any relevance in real life communication situations. The notion of 'near functional equivalence' is advantageous in that it immediately makes clear differences that are deemed important and that might have clinical value. Considering 'near functional equivalence' also allows for the flexibility of normal connected speech processes, where the influence of the surrounding sounds holds more importance than direct point-to-point comparison. A further advantage of this approach is that it can be used to measure the reliability of broad and narrow transcriptions or even a mixture of both, as comparative judgements of perceptibility can be made on any two sounds, regardless of the presence or absence of diacritics.

Future studies are needed to improve this approach. Specifically, a larger cohort of reliability checkers should be explored to decrease subjectivity. Additional studies could also determine which transcription differences could be considered 'near functional equivalence' through consensus discussions or listening activities involving phoneticians as well as SLTs.

CONCLUSION

This study has shown that measuring reliability between phonetic transcripts is not straightforward. A simple point-to-point transcription may miss the fact that some differences between transcripts represent differences which are imperceptible in everyday connected speech. Acoustic analysis provides an alternative and more objective approach to confirming transcriptions of speech samples, but to date, reports of transcriptions using data from large datasets has typically relied on perceptual methods. Moreover, if the differences between two transcriptions are ‘near functional equivalence’, the presence of a difference as observed through acoustic analysis, would still be negligible in a real-life context.

An alternative approach to measuring reliability using ‘near functional equivalence’ is provided in this report. This method is transparent in that it classifies the differences that are observed. However, it also enables a quantitative calculation of the degree to which the differences observed in pairs of transcriptions are meaningful in real life communication. In the present study, although ‘near functional equivalence’ accounted for 5.3% difference between the transcript pairs overall, of all the differences, nearly a quarter could be classed within this group.

STATEMENTS:

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We are also grateful to Joy Newbold for her work on this study and transcription of the speech samples.

Statement of Ethics

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

Disclosure Statement

The authors have no conflict of interest to declare. The authors alone are responsible for the content and writing or the paper.

Funding Sources

The UK Medical Research Council and Wellcome (Grant re: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors who will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (https://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This research was specifically funded by a National Institute of Health Research Fellowship and North Bristol NHS Trust Research Capability Funding. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the NIHR, or the Department of Health.

Author Contributions

The authors all contributed to the paper, each focusing on specific sections such as the introduction, methods and discussion. Discussions between all authors were continuous and lead to the final conclusions.

REFERENCES

[1] Müller N, Damico JS. A transcription toolkit: theoretical and clinical considerations. Clinical Linguistics and Phonetics. 2002; 16: 299-316.

[2] Ball M, Howard S, Muller N, Granese A. Data processing: Transcriptional and impressionistic methods. In: Muller N, Ball M, editors. Research Methods in clinical Linguistics and Phonetics. Chichester: Wiley-Blackwell; 2013. pp 177-194.

[3] Ball M, Müller N, Klopfenstein M, Rutter B. The importance of narrow phonetic transcription for highly unintelligible speech: Some examples. Logopedics Phoniatrics Vocology. 2009; 34: 84-90.

[4] Cucchiarini C. Assessing transcription agreement: Methodological aspects. Clinical Linguistics and Phonetics. 1996; 10: 131-155.

[5] Rutter B, Cunningham S. The recording of audio and video data. Guide to research methods. Clinical Linguistics and Phonetics. 2013; 160-76.

[6] Shriberg LD, Lof G. Reliability studies in broad and narrow phonetic transcription. Clinical Linguistics and Phonetics. 1991; 5: 225-279.

[7] Sescleifer AM, Francoisse CA, Lin AY. Systematic review: Online crowdsourcing to assess perceptual speech outcomes. Journal of Surgical Research. 2018; 232, 351-364.

[8] McAllister Byun T, Halpin P, Szeredi D. Online crowdsourcing for efficient rating of speech: A validation study. Journal of Communication Disorders. 2015; 53:70–83.

[9] Wren Y, Roulstone S, Miller LL, Emond A, Peters T. The prevalence, characteristics and risk factors of persistent speech disorder. Journal of Speech, Language and Hearing Research. 2016; 59, 647-673.

[10] Shriberg LD, Tomblin JB, McSweeny JL. Prevalence of speech delay in 6-year old children and comorbidity with language impairment. Journal of Speech, Language and Hearing Research. 1999; 42, 1461-1481.

[11] Ball M, Rahilly J. Transcribing disordered speech: the segmental and prosodic layers. Clinical Linguistics and Phonetics. 2002; 16: 329-344.

[12] Ferrier LJ, Johnston JJ, Bashir AS. A longitudinal study of the babbling and phonological development of a child with hypoglossia. Clinical Linguistics and Phonetics. 1991; 3, 187-206.

[13] Otomo DK, Stoel Gammon C. The acquisition of unrounded vowels in English. Journal of Speech and Hearing Research. 1992; 35, 604-616.

[14] Pye C, Wilcox KA, Siren KA. Refining transcriptions: The significance of transcriber 'errors'. Journal of Child Language. 1988; 15: 17-37.

[15] Bosma Smit A, Mann Brumbaugh K, Weltsch B, Hilgers M. Treatment of phonological disorder: A feasibility study with focus on outcome measures. American Journal of Speech-Language Pathology. 2018; 1-17.

[16] Viereggew H. Basic aspects of phonetic segmental transcription. In: Almeida A, Braun A, editors. Problerne der phonetischen Transkription. Franz Steiner Verlag: Wiesbaden; 1987.

[17] Shriberg LD, Kwiatkowski J, Hoffman K. A procedure for phonetic transcription by consensus. Journal of Speech Language and Hearing Research. 1984; 27: 456-465.

[18] Shriberg LD, Kent RD. Clinical phonetics (3rd Ed). Denver: Pearson; 2002. p. 372-373.

[19] Kluender L, Keifte M. Speech perception within a biologically realistic information-theoretic framework. In: Gernsbacher A, Traxler M, editors. Handbook of psycholinguistics; 2nd ed. London: Elsevier; 2006. p153-199.

[20] Howard S, Wells B, Local J. Connected speech. In: Ball MJ, Perkins MR, Muller N, Howard S, editors. The handbook of clinical linguistics. Blackwell: Oxford; 2008. p.583-602.

[21] Howard S. Phonetic transcription for speech related to cleft palate. In: Howard S, Lohmander A, editors. Cleft Palate Speech Assessment and Intervention. Chichester: Wiley-Blackwell; 2011 p.127-142.

[22] Boyd A, Golding J, MacLeod J, Lawlor D, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey-Smith G. Cohort profile: The 'Children of the 90s' – the index offspring of the

Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology. 2013; 42: 111-127.

[23] Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod, J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort profile: The Avon longitudinal study of parents and children: ALSPAC mothers cohort. International Journal of Epidemiology. 2013; 42:97-110.

[24] Paden EP, Novak MA, Beiter AL. Predictors of phonologic inadequacy in young children prone to otitis media. Journal of Speech and Hearing Disorders. 1987; 52: 232-242.

[25] Reynell J. The Reynell Developmental Language Scales - Revised Edition. Windsor; NFER-Nelson, 1977.

[26] Renfrew CE. Bus Story Test: A test of narrative speech (4th eds): Chesterfield; Winslow Press Ltd, 1997.

[27] Byrne B, Fielding-Barnsley R. Recognition of phoneme invariance by beginning readers. Reading and Writing. 1993; 5: 315-324.