Published in final edited form as:

Clin Linguist Phon. 2007 October ; 21(10): 793–831. doi:10.1080/02699200701547869.

Predicting phonetic transcription agreement: Insights from research in infant vocalizations

HEATHER L. RAMSDELL, D. KIMBROUGH OLLER, and CORINNA A. ETHINGTON

The University of Memphis, TN, USA

Abstract

The purpose of this study is to provide new perspectives on correlates of phonetic transcription agreement. Our research focuses on phonetic transcription and coding of infant vocalizations. The findings are presumed to be broadly applicable to other difficult cases of transcription, such as found in severe disorders of speech, which similarly result in low reliability for a variety of reasons. We evaluated the predictiveness of two factors not previously documented in the literature as influencing transcription agreement: canonicity and coder confidence. Transcribers coded samples of infant vocalizations, judging both canonicity and confidence. Correlation results showed that canonicity and confidence were strongly related to agreement levels, and regression results showed that canonicity and confidence both contributed significantly to explanation of variance. Specifically, the results suggest that canonicity plays a major role in transcription agreement when utterances involve supraglottal articulation, with coder confidence offering additional power in predicting transcription agreement.

Keywords

Phonetic transcription reliability/agreement; canonical babbling; transcriber confidence

Introduction

Introductory rationale: The role of canonicity in phonetic transcription

Study of infant vocalizations through transcription is inherently complicated due to the immaturity of infant sounds. Researchers attempting to characterize immature sounds phonetically often “shoehorn” (Oller, 2000a) them into an adult model through phonetic transcription. Clearly it is important to consider ways such transcription poses challenges to inter-observer agreement or reliability.¹

The present article offers an empirical perspective on inter-transcriber agreement in difficult cases such as those encountered in research on infant vocalization. We cast the research against a theoretical framework that clarifies fundamental reasons for problems of agreement in transcription of immature sounds. The framework has been developing in the field of infant vocalizations for over 30 years (see Oller, Wieman, Doyle, & Ross, 1975; Stark, 1978; Koopmans-van Beinum & van der Stelt, 1986), and specifies infrastructural requirements of well-formedness or *canonicity* of mature speech sounds (Oller, 1980). The immaturity of infant

© 2007 Informa UK Ltd

Correspondence: Heather L. Ramsdell, The University of Memphis, School of Audiology and Speech-Language Pathology, 807 Jefferson Avenue, Memphis, TN 38105, USA. hramsdll@memphis.edu.

¹Throughout this paper the terms “transcription reliability” and “transcription agreement” are used interchangeably, although “agreement” is, in principle, a type of “reliability”. Cucchiari (1996) clarifies the distinction.

sounds can be elucidated within this framework. In short, vocalizations of infants in the first half year of life usually fail to meet the infrastructural standards of canonicity, and, thus, to varying degrees across utterances, fail to conform to principles known to characterize the overwhelming majority of canonical syllables in mature spoken languages all over the world.

A primary characteristic differentiating infant from mature vocalizations concerns articulatory transitions. Typically developing infants and speakers with communication disorders (e.g. with deafness or dysarthria) of any age often produce slow articulations from consonants (C) to vowels (V), a pattern that results in slow formant transitions (Oller, 1980). Listeners recognize syllables with transitions that are slower than those typically found in mature speech, and report that the syllables are drawn out, slurred, fuzzy, indistinct, or distorted. Such syllables may also include low slope in their formant transitions, a pattern that has been reported (especially for F2) to characterize relatively unintelligible speech produced by individuals with dysarthria (Weismer, 1984; Kent et al., 1989; Weismer & Martin, 1992). Thus, slowing of articulatory transitions and consequent reduction in slope of formant transitions can lead to violation of the rapid transition principle for canonical syllables, one of the critical infrastructural properties that characterizes well-formed syllables in natural languages.²

When syllables violate universal principles of canonical syllable formation in spoken languages, it stands to reason they may be hard to identify because they fail to conform to the syllabic templates that mature listeners use as reference standards in perception. The notion of templates or prototypes for categorization and recognition is widely utilized in perceptual theory (Kuhl, 1992; Goldstone, 1998; Barrett, 1999; Lu & Doshier, 2004). The basic idea is that perception and perceptual learning are organized around abstract models (or templates) of perceptual categories. Templates are stabilized through experience with exemplars of categories. Variation in presented stimuli from well-established templates results in perceptual variability. Canonical syllables, because they are so common in speech, provide the basis for listeners to form a template against which vocalizations with varying degrees of syllabic organization can be judged for well-formedness. Further, individual canonical syllable types that occur relatively often in languages can be predicted to be relatively easily identified

²The criterion for judgement of canonicity in infant vocalizations has always been (and to the present remains) primarily auditory. The methods section provides a description of the auditory judgement procedure that has been used in the second author's laboratories for years. A primary reason that we continue to focus on auditory rather than instrumental acoustic judgements in this research is that the definition of canonical syllables in acoustic terms is still not yet fully established. We here offer a brief summary of the acoustic status of the definition.

A criterion duration for formant transitions has been nominally established based upon acoustic examination of relatively measurable formant transitions in auditorily judged canonical and non-canonical syllables from infants, as summarized in Oller (2000a). Measurement of formant transitions in infant vocalizations can be extremely difficult, especially because high pitch of many infant syllables produces harmonics that are very widely spread. The nominal criterion based on relatively measurable transitions is 120 ms (usually focusing on F2) as a maximum for canonical syllables. This value is primarily based on infant syllables where both F1 and F2 have been reliably visible in spectrographic displays with at least 600 Hz analysis bandwidth, and where F1 and F2 vary from a consonantal locus to a nuclear (vowel) locus and then reverse slope. The end of the formant transitions can thus be referenced to a steady state or a reversal of slope.

However, beyond the nominal durational criterion, it is clear that to attain a generally applicable acoustic definition of canonical syllables, additional specification is needed to account for differing types of syllables and differing utterance-level patterns. For example, syllables with nasal or aspirated consonants often show extremely short formant transitions in acoustic displays, and we are investigating the utility of amplitude rise time as a possible substitute for or supplement to formant transition duration as a criterion for canonicity in such cases. Also, at slow speaking rates the maximum transition duration may need to be higher than at rapid rates.

A transition slope criterion is also obviously required (because if slope is too low, change in formant frequency would be heard as no change). The slope data on dysarthric patients from Kent et al. (1989) focused on circumstances where F2 appeared to be a useful focus for determining a criterion for intelligible syllables: If average F2 transition slopes were lower than a ratio of 2.5 (Hz ms⁻¹), speakers proved to be highly unintelligible. However, there are of course canonical syllable types where F2 slope is inherently low, i.e. if F2 locus for a consonant is near the F2 target for its adjacent vowel. So slopes of other formants (F1 and/or F3) may need to be referenced to determine canonicity in such cases. Further, the slope criteria for canonical transitions suggested by the adult F2 data would presumably need to be normalized for infant formant values which are known to vary widely from those of adults. As research proceeds towards the development of a more elaborate and finely tuned acoustic definition of the notion canonical syllable, it will have to make reference to a wide variety of acoustic facts, but the success of the approach will always need to be referenced to auditory judgements of real listeners about well-formedness. In the meantime, auditory judgements remain at centre stage in the judgement of canonicity.

because each such canonical syllable corresponds to an accessible template for experienced listeners. Variation in acoustic signals from those templates, as would occur with all non-canonical syllables, should yield variable perception, and, according to our reasoning, variable transcription. The template theory suggests a fundamental role for familiarity of stimuli in recognition.

Many other researchers have recognized infrastructural aberrations in non-canonical infant sounds, and have taken note of their troublesome effects on perception and transcription. The aberrations are so disruptive to description in traditional phonetic terms that Bauer (1988), for example, proposed a shift of terminology—he suggested the terms “consonant and vowel” be changed to “closant and vocant” for infant sounds in order to highlight the fact that infant vocalizations often violate the canonical standards for consonant and vowel characteristics in the syllables of mature languages. The present authors utilize the traditional terms “consonant-like” and “vowel-like” when we refer to pre-canonical infant vocalizations for the same reason—the closures and openings of the vocal tract during infant vocalizations are hard to identify as pertaining to particular Cs or Vs, precisely because they vary infrastructurally from well-formed Cs and Vs composing canonical syllables.

In view of the problems of auditorily identifying segments in infant vocalizations, researchers have largely abandoned the assumption that infants in the first year of life command the production of independent Cs and Vs. They postulate instead that infants control at most a *whole-syllable* level of production (MacNeilage & Davis, 1990; Davis & MacNeilage, 1995). However, even these postulated whole syllables produced by infants typically lack full canonicity (Oller, 1980; Stark, 1980), often due to excessively long (slow) articulatory transitions.

Slow transitions disrupt the rhythmicity of speech, and have been proposed as a major factor in the breakdown of reliable auditory identification of syllables. Indeed, early infant vocalizations have been reported to produce significant agreement problems between transcribers (Lynip, 1951; Duckworth, Allen, Hardcastle, & Ball, 1990). Lynip went so far as to suggest that phonetic transcription should be abandoned entirely for infant sounds partly because of inter-transcriber agreement problems. Others have taken a more moderate stand, restricting samples that are to be transcribed to utterances deemed canonical (therefore meeting the infrastructural requirements of mature speech), and consequently focusing on vocalizations from infants who have already entered the canonical stage (e.g. Oller et al., 1975; MacNeilage & Davis, 1990, 2000; Davis and MacNeilage, 1995).

The idea that low canonicity contributes to problems of reliable identification of speech sounds (and consequently to problems of transcription reliability) has become a sort of common wisdom in research on infant vocalizations, even though there is no written publication to our knowledge providing empirical demonstration of the presumed effect of canonicity on transcription reliability. As far as we know, the only formal report on the topic thus far was presented orally at the International Child Phonology Conference by the second author, who offered preliminary data suggesting that non-canonical infant utterances yield lower transcriber agreement than canonical utterances produced by infants of a similar age (Oller, 2000b). The primary goal here is to address more formally canonicity as a potential source of variance in transcription reliability for difficult cases such as those presented by infant vocalizations.

Reliability and validity

Lack of full canonicity in phonetically transcribed vocalizations also presents a challenge in terms of *construct validity*. The generally accepted definition of construct validity has been encapsulated by Kirk (1978: 108): “The validity of a test is the degree to which it measures what it is supposed to measure”. Transcription of infant sounds may run afoul of construct

validity in the sense that the transcribed symbols may simply *not* correspond to the sounds actually contained in the vocalizations. If vocalizations include sounds that differ from mature speech sounds to varying extents, then transcription provides a tool that will fail to measure what it is supposed to measure to varying extents depending on the degree to which the vocalizations in the sample fail to conform to the infrastructural principles required to build mature sounds of speech.

An implication of limitations in construct validity is that phonetic transcription for many infant sounds cannot, strictly speaking, be judged in terms of *accuracy*. Of course, transcriber agreement provides no guarantee of transcription accuracy, even with mature speech. Two transcribers can agree and both simply be wrong, a fact that can be verified whenever a *gold standard* transcription can be externally validated. However, in the case of pre-canonical vocalizations of infants, the inherent ambiguity of the signal often precludes such external validation by a gold standard transcription. Immature, pre-canonical vocalizations routinely *do not admit of a single correct transcription*. The sound sequences commonly vary so much from canonicity that multiple phonetic interpretations prove possible, and there is often no way to resolve them by consensus. Even individual highly trained listeners working with infant vocalizations often experience the well-known psychological phenomenon of perceptual set (Gibson, 1941; Epstein & Rock, 1960; Ralston & Johnson, 1990), wherein an utterance can be perceived in different ways on different occasions as the observer specifically focuses on the utterance with different expectations.

A sense of ambiguity about the IPA interpretation for infant sounds has been universally reported by many highly trained individuals, including a number of experienced academic phoneticians who have attempted such transcription in the laboratories and collaborating laboratories of the second author over the past 30 years. Research on perceptual illusions and variability of perception has provided empirical confirmation of the experience that speech-like and other complex sounds can be multiply interpreted (Warren & Obusek, 1971; Deutsch, 1997), and may be influenced by contextual information (McGurk & MacDonald, 1976) and phonetic expectation (Oller & Eilers, 1975).

Rather than assuming that a “correct” transcription exists against which accuracy can be judged in such cases, we deem it more appropriate to accept that variability in judgement is a part of the game—a reflection of inherent ambiguity in the infant signal with respect to the coding system. No individual transcription is assumed to be correct; instead, a set of independently performed transcriptions by trained listeners is assumed to provide an inventory of plausible interpretations. At the same time, of course, differences among these transcriptions are also indicators of inter-transcriber disagreement. Both the range of plausible transcriptions and the inter-transcriber agreement are affected by the maturity and disorder status of the speaker. Research has illustrated this point empirically (Shriberg & Lof, 1991; Stoel-Gammon, 2001) in the sense that transcriptions of unimpaired adult speakers tend to yield higher agreement and a smaller range of alternative transcriptions than those of young children or speakers with impairments.

With this perspective on canonicity and its role in perception in mind, the research reported here provides a special view on construct validity for phonetic transcription. We suggest that the degree to which sounds abide by principles of canonicity provides a gauge of the extent to which transcription could be thought of as a valid description for those sounds. A primary hypothesis being investigated here is that inter-observer reliability should decline as canonicity declines, which implies that reliability is likely to drop as construct validity of the measure (i.e. transcription) drops. This is not to suggest that reliability and validity are the same concepts, but merely that, in this case, the level of reliability is expected to covary with validity. The present study draws on the infra-phonological framework, wherein the construct validity of

transcription (as seen through the lens of canonicity) provides a predictive backdrop against which important aspects of transcription reliability can be evaluated and interpreted (Oller, 2000a).

Confidence judgements and transcription

In the present article, we introduce coder confidence as an additional potentially important factor in the study of transcription agreement. In fact, coder confidence has influenced research on infant vocalizations in the laboratories of the second author and his collaborators since the early 1970s. Even the formulation of the concept of canonicity was influenced by coder reports of lack of confidence in the transcription of sounds judged to be pre-canonical. We have long taken these reports seriously because they reflect natural auditory reactions to developing speech control in infancy. When parents and other caretakers listen to infant vocalizations, they appear to listen for the first occurrences of controlled speech-like sounds that might be deemed mature enough (that is, canonical enough) to be used in speech. With the onset of canonical babbling (usually by early in the second half year of life), parents (at least in middle SES European and American families) systematically accelerate active word teaching, seeking meanings to associate with their canonical child's utterances (Papoušek, 1994). Parents must be capable of recognizing canonical babbling or else they could not determine the appropriate point in time to accelerate this intuitive instructional process. Consequently, they must be able to differentiate between pre-canonical and canonical sounds. In fact, research has indicated that untrained parents are quite accurate in describing their infant's vocalizations. On the basis of parent descriptions, laboratory staff are able to determine whether an infant has reached the canonical stage, and this fact can be verified by other laboratory staff who interact with the infant but do not know how the parent described the infant's sounds (Oller, Eilers, & Basinger, 2001).

Indeed, parents tend to express *confidence* that their infants can produce certain canonical patterns (e.g. [dada] or [baba]) once the stage has begun. And, like laboratory staff, their confidence level in identification of particular syllables appears to be an indicator of their awareness of the canonicity of those syllables. Thus, confidence levels of observers have played an important role in the development of the infra-phonological model with its key concept of canonicity.

Additionally, confidence levels of observers are widely utilized to enhance understanding of perceptual judgements in a variety of fields. Baranski and Petrusic (1998: 929) point out that: "There has been a fascination with the study of confidence in human judgement since the beginning of experimental psychology". Confidence of coders on individual judgements is a key principle in signal detection theory (Swets, 1964; Green & Swets, 1966), having been shown often to be highly correlated with receiver operating characteristics (ROC curves) (Pollack & Decker, 1964; Baranski & Petrusic, 1998; Van Zandt, 2000). Confidence judgements have been shown to be related to both accuracy of judgement and reliability on individually coded events (Pollack and Decker, 1964), and such research has played important roles in, for example, the interpretation of eyewitness testimony (Wells & Murray, 1984; Robinson & Johnson, 1996; Roberts & Higham, 2002) and medical diagnosis (Richards, Hicken, Putzke, Ness, & Kezar, 2002; Marbach, Franks, Raphael, Janal, & Hirschhorn-Roth, 2003).

Yet, there has been no written published report on confidence levels regarding canonicity to our knowledge. The reports of listeners indicating lack of confidence in recognition and transcription of pre-canonical utterances are anecdotal, and there are no published quantitative data to indicate the degree of association between canonicity and confidence in transcription. The only relevant information is the second author's (Oller, 2000b) previously mentioned oral

report, which also provided preliminary evidence that canonical utterances yield higher transcriber confidence than non-canonical utterances.

Similarly, as far as we can tell, there are no published reports on confidence levels of observers in relation to phonetic transcription agreement. Given the widespread use of confidence levels for interpretive enhancement of perceptual research in other domains, the absence of confidence research in phonetic transcription is salient.

Background on transcription agreement

Having claimed that both canonicity and confidence levels have been ignored in prior research on transcription agreement, we hasten to acknowledge that research on other aspects of transcription agreement has been substantial. Shriberg and Lof (1991) reviewed existing literature on transcription agreement, and provided extensive empirical data from their own laboratory. An interesting feature of the research literature is the indication of extremely variable levels of overall agreement across samples of vocalization depending on a variety of factors. Shriberg and Lof reported levels of agreement from well over 90% to as low as 20%, depending on sample characteristics and other procedural variables. We have similarly found levels of overall agreement ranging from 96% for transcription of words produced in citation form by a native American English speaking adult to 18% for transcription of utterances from the sample of infant vocalizations studied here (Oller & Ramsdell, 2006).

Shriberg and Lof (1991) posited 16 variables that contribute to transcription reliability. Some of the variables pertain to characteristics of *subjects* providing samples to be transcribed, some to *analysis* procedures and transcriber characteristics, some to *contexts* of sampling, and some to the particular phonetic *units* that are selected as a focus for analysis. All of these variables have empirical support (Shriberg & Lof, 1991; Louko & Edwards, 2001; Stoel-Gammon, 2001; Coussé, Gillis, Kloots, & Swerts, 2004). In response to all the sources of variance and the extreme discrepancies in reliability outcomes from the many studies they reviewed, Shriberg and Lof (1991: 230) concluded "... the diversity of interactive factors ... makes it difficult to formulate generalizations about the reliability of phonetic transcription". A review by Stoel-Gammon (2001) amplified the discussion of sources of variance in transcription agreement, with particular focus on age of subjects and the extent to which samples included babbling as opposed to real speech.

It is notable that canonicity and coder confidence simply do not appear as factors in this literature and are not addressed in either the review by Shriberg and Lof or that by Stoel-Gammon. Yet these factors may correlate with and offer important additional perspectives on transcription agreement and on their role with respect to other factors that have been previously posited to affect transcription agreement.

Goals of the research

Correlational analysis was applied to transcription agreement and confidence judgement data from coding of infant vocalizations to assess two key relations that have been long presumed to be significant but never before quantitatively assessed. The goals of these correlational analyses were:

- (1) To determine whether utterance canonicity is significantly related to transcription agreement, and
- (2) To determine whether transcription confidence is significantly related to utterance canonicity.

Goals 1 and 2 target zero-order correlations to assess predictions of the present study. In addition, however, the design of the study allowed comparisons of the *relative* contributions

of the canonicity and confidence variables in prediction of transcriber agreement. Regression analysis was applied:

- (3) To evaluate the relative roles of utterance canonicity and coder confidence in transcription agreement as reflected by the proportion of variance in transcription agreement explained by each, and to determine whether confidence judgements account for reliable additional variance in transcription agreement beyond that accounted for by canonicity.

Beyond the study's primary goals, which targeted canonicity and confidence variables as correlates of transcription agreement, the naturalistically acquired utterances also provided the opportunity to view the role of canonicity in the context of additional factors posited to affect transcription agreement. First, each of the primary vocalization sample's utterances were categorized as canonical or non-canonical. The difference between transcription agreement for the canonical and non-canonical utterances was then assessed by t-test to provide a basis for comparison with the additional factors. The two additional factors that were evaluated by t-tests through the data corpus were non-nativeness of transcribed segments and vocal quality aberrations of utterances. Both of these additional factors, as well as canonicity, can be thought of as related to familiarity of signals. In each case, the factor to be evaluated can be thought of as pertaining to whether (or to what extent) stimuli for transcription match the templates that listeners presumably utilize as reference points in identification.

The non-nativeness factor compared agreement on transcribed IPA segments not occurring in standard American English phonetics to transcribed segments that do occur in standard American English (including all its standard allophones and free variants), the primary language of all the transcribers. The analysis took advantage of the fact that the primary vocalization sample often included sounds that were transcribed as not pertaining to the English repertoire. If the standard transcriber indicated a segment as non-native, it was predicted that the segment would yield lower transcription agreement with the comparator transcribers than if the segment was transcribed as one of the American English elements.

The vocal quality analysis took advantage of the fact that some of the utterances in the sample included substantial deviations from modal voice, such as falsetto, pressed voice, or creakiness. It was predicted that utterance-level transcription agreement would be lower in the presence of vocal quality deviations than in their absence. Both non-nativeness and vocal quality aberrations presumably represent deviations from standard perceptual templates, and of course, lack of canonicity can also be viewed as such a deviation. Consequently a t-test analysis was provided:

- (4) To assess the roles of canonicity, non-nativeness, and vocal quality aberrations in transcription agreement.

The data also provided the opportunity for item analysis on transcription agreement in the study. We tabulated common transcription disagreements and provided a comparison of the degree of transcription agreement for consonants vs vowels:

- (5) To provide an item analysis on disagreement patterns in transcription of infant vocalizations.

Finally, as an empirical supplement to the study, analysis of transcription reliability for additional coders on utterances drawn from three additional infants was conducted. The supplementary sample provided tighter control over variables so that a targeted comparison of transcription agreement in canonical and non-canonical utterances could be made. The key goal was:

- (6) To assess by t-test the generalizability to additional coders and infants of the relation of canonicity and coder confidence to transcription reliability.

Method

Coding sample

Primary vocalization sample—The primary vocalization sample analysed here was used in a prior study (Oller & Ramsdell, 2006), which laid the groundwork for this investigation. Thirty infant utterances were coded by 8 listeners at 6 levels (3 involving utterance judgements, and 3 involving confidence judgements). The codings were paired and aligned at the segment (nucleus or margin) level. One coder was selected as the standard comparator for the 7 coder pairings. Thereafter, an automated analysis was conducted to determine transcription agreement and a variety of additional analyses were conducted to relate transcription agreement with canonicity judgements and confidence judgements.

The 30 utterances were judged to include over 100 segments by all the coders, which resulted in an average of 117 comparisons (at the segment level) for each of the 7 coder pairings (each of 7 coders paired with the standard coder) at each of the 6 levels of coding, leading to 4914 ($117 \times 7 \times 6$) comparisons of coder judgements in the initial analysis of the data. In each of the additional analyses (most focusing on a single level of coding), the number of comparison points (the total n) equals or exceeds 700. This sample size was apparently adequate to our goals, since the results reported below are robust, providing many statistically reliable effects. The limitation in number of utterances in the present work was intentional, given the complexity of the analysis to be conducted and the interest in assessing targeted relations with precision.

The utterances for the primary analysis were carefully selected to be of varying lengths, each one consisting of a single breath group with continuous phonation except for breaks perceived as pertaining to consonant-like elements. They were extracted from two recording sessions of a longitudinal study with a normally developing infant at the beginning of the canonical stage of vocalization. The sessions were recorded on high fidelity audio equipment, digitized, and imported into LIPP™ (Logical International Phonetics Programs), which is a computer program that allows for transcription in IPA via the traditional keyboard (Oller & Delgado, 1999).

The vocalizations were chosen from sessions including both canonical and pre-canonical utterances in an attempt to represent the range of sounds the infant was producing at that stage. Based on the second author's coding, ~40% of the syllables were canonical. In addition, the utterances included examples of squeals and isolated vowel-like elements, as well as marginal syllables (with non-canonical transitions). Vocalizations were extracted from sessions during which the infant was quite vocal, with considerable parent–infant interaction. The selected utterances were thus presumed to be exemplary of a range of vocalizations in the infant's repertoire at the time of the recording. The digitized utterances, clipped out of the samples so that only the infant's voice could be heard, showed good signal quality through 10 kHz.

Supplementary vocalization sample—The principal focus of this article was based on analysis of the primary sample, but it was deemed instructive to conduct analyses on a supplementary vocalization sample from additional infants to enhance the generalizability and clarity of the results. The goal was to provide a t-test comparison of transcription agreement for utterances composed of canonical syllables as opposed to utterances composed substantially of non-canonical syllables. The supplementary sample, drawn from on-going longitudinal research in our laboratories, consisted of vocalizations from three additional infants during the first year of life, with both canonical and non-canonical utterances of high signal quality selected from each of the infants.

While the primary sample was selected to be representative of *all* pre-canonical vocalizations thought to be precursors to speech in infancy, the supplementary sample was specifically restricted in order to target canonicity more specifically. First, the supplementary sample included only utterances that had consonant-like articulations, so that canonicity of articulatory transitions could be sharply contrasted in designated canonical and non-canonical utterances. Secondly, the supplementary sample included only utterances with no substantial vocal quality aberrations, a restriction that precluded any confound of vocal quality effects in the comparison of transcription agreement and coder confidence for canonical vs non-canonical utterances. Thirdly, the supplementary sample was selected such that sounds transcribed as non-native with respect to the American English inventory of allophones and free variants were concentrated in the canonical utterances. The second author's transcriptions were utilized to verify the assignment. There were 18 segments transcribed as not pertaining to the English phonetic repertoire (non-native sounds for the coders) in the canonical utterances and only 4 in the non-canonical utterances. This precaution was taken to avoid the possibility that higher transcription reliability on canonical utterances could be attributable to a greater number of non-native sounds in the non-canonical utterances.

Twenty utterances meeting the criteria were selected, 10 of which were categorized as primarily canonical and 10 as primarily non-canonical by the second author. The second author's transcriptions of these vocalizations indicated that 35 of 37 syllables met the canonicity standard for articulatory transitions among the canonical utterances, while only 4 of 34 syllables met the canonicity standard in the non-canonical utterances. Each designated non-canonical utterance had at least one non-canonical transition corresponding to a marginal syllable according to the definitions of Oller (2000a), and there were 17 designated non-canonical transitions among the 10 non-canonical utterances. A single non-canonical transition was perceived to occur in 2 designated canonical utterances, but each of those utterances included 3 additional designated canonical transitions. As a consequence, the supplementary sample provided a sharper distinction between canonical and non-canonical utterances in terms of the articulatory transition criterion than the primary sample.

These 20 utterances were randomized and transcribed by 4 coders. The transcriptions yielded an average of 128 comparisons (at the segment level) for each of the 3 coder pairings (i.e. each coder paired with the standard) at each of the 3 levels of coding, leading to ~1152 (128×3×3) comparisons of coder judgements in the analysis.

Coders

Primary sample—The 8 coders for the primary vocalization sample were students and faculty in the School of Audiology and Speech-Language Pathology at the University of Memphis. Four were master's students and 2 doctoral students (one of whom is the first author) who had previously been trained intensively in a phonetic transcription class presented by the second author. The training utilized the IPA as implemented in the LIPP™ software program. The other 2 transcribers were doctoral faculty with expertise in phonetics. We chose to work with 8 observers to provide a perspective on the range of transcriptions that may occur during infant vocalization coding, and in the hope of providing greater generalizability of our findings across potential coders. At the same time, it should be noted that generalizability of the results is limited by the facts that the coders were all native English speakers and the students were all trained in phonetic transcription by the second author.

The second author is one of the doctoral faculty, an academic phonetician for over 30 years and the most experienced transcriber of the group. He is also a proficient speaker of 3 European languages in addition to English, making him the most experienced polyglot of the group. Further, he is the primary developer of the infra-phonological model upon which the canonicity

definitions are based. For these reasons, the second author's transcriptions were used as comparator (or standard) codes for the analyses.

Even so, there is no assumption in this work that the comparator coder's transcriptions were "correct", given the above reasoning regarding inherent ambiguity in phonetic interpretation of infant utterances. Furthermore, when a second coder from the group was chosen as the standard, and the results were recomputed, the overall transcription reliability changed extremely little—in fact, the outcomes were identical when rounded to 2 significant digits (.60 proportion of agreement in the case of both standard transcribers—see below for explication of the metric of agreement). Additionally, the overall transcription agreement results were quite similar for master's students (.60 agreement with the standard), doctoral students (.58), and faculty (.62). A one-way analysis of variance tested the variation in the weighted reliability values for the 3 groups, yielding non-significant results [$F(3, 116) = .361, p = .781$]. Consequently, in the results reported below, the data from each of the 7 comparator coders were treated equivalently (not grouped) in the analyses.

Differences in agreement among coders or groups of coders were very small in comparison with differences in agreement across the utterances in this study—an indication that we succeeded in selecting utterances of a wide range of transcribability. As reported in our prior publication using this sample (Oller & Ramsdell, 2006), mean reliability for all the coder pairings ranged from .85 to .33 across the 30 utterances. Thus, the difference in average transcription reliability across utterances (.85–.33=.52) was 13-times greater than the maximum difference between coder groups (.62–.58=.04).

Supplementary sample—Two of the 4 coders for the supplementary sample overlapped with the primary sample coders; they were the first and second authors, a doctoral student and a faculty member. Additionally, 2 coders for the supplementary sample were master's students who had been trained to a high standard of IPA coding in the context of other research utilizing phonetic transcription under the direction of the second author. These 2 coders had spent more than a year in weekly supervised phonetic transcription of child speech samples. Both coders were native speakers of American English. For the supplementary analysis there were 3 coder comparisons, again with the second author treated as the standard.

Coding design and protocol

Primary sample: Canonicity judgements—The coding of canonicity is founded on the infra-phonological model and its definition of canonicity (Oller, 2000a). Coding was conducted at a segmental level, with each C (margin or consonant-like element) and N (nucleus or vowel-like element) designated by transcribers at the first step of coding on each utterance. An N was defined as a syllable nucleus and, thus, the number of syllables in each utterance, as perceived by each transcriber, was the number of Ns assigned by that transcriber. The C and N segmentation provides a frame around which we can explicate two principles that the transcribers were required to employ in their judgements of canonicity.

The canonical transition principle—According to definitions of the model detailed in Oller (2000a), a canonical syllable must include at least one N and at least one supraglottally articulated C. The transition principle of canonical syllables (also discussed above) focuses on the perceived temporal relation between C and N. The rule of thumb that the transcribers were instructed to use to identify canonical and non-canonical transitions was consistent with that employed in much prior research designating canonical syllables (e.g. Oller, 1980; Lynch, Oller, Steffens, & Levine, 1995; Nathani, Ertmer, & Stark, 2006). The rule was that if transitions could be perceived *as transitions* auditorily, they were to be categorized as non-canonical. Canonical transitions, on the other hand, are perceived as gestalt syllable patterns,

such that the movement from C to N is not itself isolable auditorily from the whole. Canonical syllables with stop consonants (e.g. [ba], [bi]) exemplify this gestalt pattern. If the transitions between C and N in such syllables are slowed progressively through speech synthesis, then perception of the syllables progressively changes to semivowels (e.g. [wa], [wi]); with further lengthening, the gestalt syllable sense is lost, and the outcome is perceived as non-canonical or as multisyllabic (e.g. [ua], [ui]) (Liberman, Delattre, Gerstman, & Cooper, 1956). Based on the formant transition durations typically associated with glide consonants such as [w], the limit on canonical articulatory transition durations has been nominally set at a 120 ms maximum (Oller, 1980), although the appropriate value appears to be sensitive to a variety of factors such as the duration of adjacent segments. However, the judgements of the coders in this study were made auditorily, according to the procedure indicated above (see note ²).

Figure 1 spectrographically illustrates the distinction in transition duration between a canonical and a non-canonical syllable from the primary infant vocalization sample used in this study. The syllables were clipped from two different utterances and pasted side-by-side for the illustration. As can be seen, the releasing formant transitions in the canonical syllable are relatively quick (68 ms), while those in the non-canonical syllable (147 ms) exceed the 120 ms nominal maximum. In accord with the expectations of the infra-phonological model, the canonical syllable is heard as a gestalt, while the non-canonical syllable is heard in such a way that the transition from C to V can be tracked auditorily in real time—the C and V can thus be heard as individual segments with a gradual change between them. In accord with our expectation, the transcription agreement for the canonical syllable was much higher than for the non-canonical syllable (see results).

In the coding of the primary sample, the transcribers placed an asterisk indicating non-canonical status in the column corresponding to any C where a non-canonical transition to an adjacent N occurred. Either NC or CN sequences could be designated as non-canonical. NCN sequences could be non-canonical either due to arresting or releasing transitions violating the transition principle.

The canonical nucleus principle—Typical vowels occurring in mature speech can be termed canonical nuclei. Infants, however, often fail to produce nuclei in the way mature speakers do. In particular, infants often produce *quasivowels*, which are by definition non-canonical. A quasivowel is a vowel-like sound produced with the vocal tract *at rest*; in contrast, a vowel sound produced with a *postured* vocal tract is categorized as a canonical nucleus or *full vowel* (Oller, 1980). Accordingly, a quasivowel cannot be the N of a canonical syllable. Usually quasivowels tend to sound like high central unrounded nasalized vowels or isolated (syllabic) nasal consonants, and tend to show a spectral tilt with primarily low frequency energy. The key feature for auditory identification of quasivowels is the impression of phonation occurring without any intentional movement or posturing of the supraglottal vocal tract (Oller, Eilers, Bull, & Carney, 1985). Transcribers in the present study indicated that an N was a quasivowel by placing an asterisk in the column associated with each N that they coded.

Just as syllables with non-canonical transitions tend to occur more frequently early in development, so do quasivowels tend to be produced early in development and to fall off as a proportion of all utterances across the first year of life. In contrast, full vowels increase in proportion across the first year, a pattern that corresponds to the growing control of the infant over vowel-like sounds produced with varying articulatory postures and resonance characteristics (Oller, Eilers, Steffens, Lynch, & Urbano, 1994; Nathani et al., 2006).

Non-canonical transitions as a consequence of quasivowel nuclei—As noted above, a quasivowel cannot, by definition, participate in the formation of a canonical syllable.

In cases where a quasivowel occurs with an adjacent C, the infra-phonological framework designates the transition between the segments as non-canonical, because a canonical transition requires supraglottal articulatory movement of substantial magnitude between C and N. Very little movement occurs from a closure for a C if the target N is a quasivowel. On listening to a syllable consisting of a C followed by a quasivowel, one has the sense that the syllable is incomplete, or has been stunted in some way, as if the N were simply never fully articulated. Syllables consisting of quasivowels plus a C-like closure (either arresting or releasing) are extremely common in infancy, and are sometimes perceived as “grunts” or “little hums” with sudden supraglottally articulated onsets or offsets. These quasivowel syllables are clearly not canonical syllables to the ear because they are perceived as lacking the typical articulatory movement from C to N that characterizes the great majority of syllables in mature speech.

Such syllables often have low slope of formant transitions (i.e. little change of formant loci from C to N) compared to highly intelligible syllables, as indicated in Kent and Weismer’s formulation (Weismer, 1984; Kent et al., 1989). Non-canonical transitions, then, can come in two forms. When quasivowels are adjacent to Cs, the transition requirement of canonical syllables is violated because there is too little articulatory movement to meet the canonical transition requirement. In the case of slow transitions from Cs to full vowels, the transition requirement is violated because of the protracted duration of the transitions. In both cases, the transitions have a low degree of slope (see note ²).

Whenever transcribers indicated an N was non-canonical, any adjacent C was also automatically interpreted as non-canonical. At the point of analysis, an asterisk was inserted for such a C if it had accidentally not been recorded by the transcriber.

Protocol sequence for all judgements—Prior to initiation of the coding for the primary sample, all coders met and were presented with the coding protocol, including the canonicity definitions illustrated with auditory examples. Practice infant vocalizations were coded as a group to ensure understanding of procedure and to provide opportunity for coders to explore difficulties with transcription of infant utterances before being presented with the test sample. Explicit instruction was also given to encourage the use of IPA symbols representing sounds from a variety of languages at the point of transcription. We hoped to discourage (to whatever extent possible) bias resulting from familiarity with English phonetics, a point that had been previously emphasized in classroom training on phonetic transcription (nevertheless, in the Additional Results section it is suggested that familiarity bias was present despite these instructions). For additional practice using the protocol, a 10-utterance pilot sample, based on additional utterances from the same infant, was transcribed by all the coders and reviewed to ensure clear understanding of the procedure.

Then, the formal coding for the primary sample began, with everyone transcribing the 30-utterance experimental sample. It was required that each transcriber work completely independently and listen to the individual utterances a limited number of times, only twice at each of 6 levels of coding (a total of 12 times per utterance). The 6 steps were of 2 types, one for coding and a second for confidence judgements associated with the coding steps, represented in *italics* below (see Table I).

The confidence judgements were based on a 7-point Likert scale, with “1” representing little confidence in the judgement and “7” representing complete confidence. Our choice of 7 points is close to that found in other studies on confidence in auditory-based coding (Pollack & Decker, 1964), although a range from 3–9 points is common in research using Likert scales (Wells & Murray, 1984; Aronson, Ellsworth, Carlsmith, & Gonzales, 1990; Robinson & Johnson, 1996; Baranski & Petrusic, 1998; Roberts & Higham, 2002). We sought to ensure

that the number of points on the scale would be large enough to potentially support stable correlations with other measures, while not being so large as to confuse coders.

During the coding, no acoustic displays were employed. Coders were instructed to work in the following ordered steps, completing all steps on each utterance before proceeding to the next (see Table I for an example coded utterance):

1. record the syllable structure of the utterance by indicating all margins (Cs or consonant-like elements) and nuclei (Ns or vowel-like elements);
2. for each C and N, make a confidence judgement on the 1–7 scale regarding the existence of each margin and nucleus coded (i.e. how confident are you that each C or N you coded is present in the utterance?);
3. indicate the canonicity of each C or N by marking an asterisk in the column for each non-canonical C or N; leave canonical slots blank;
4. make confidence judgements (i.e. 1–7) on the canonicity judgements for each C or N;
5. transcribe each utterance in LIPP™ using IPA (the number of transcribed segments was required to correspond one-to-one with C and N codings from step 1); and
6. make confidence judgements (i.e. 1–7) on each IPA segment coded.

The 6 levels were implemented with the intention of differentiating aspects of the coding process. For example, listeners began in step (1) by characterizing the global structural segments (Cs and Ns) of each utterance. In step (5) they characterized the designated segments more fully with the IPA. Coders were thus forced to address syllable/segment structure before addressing details regarding components of the syllable. By having them focus first on global syllable/segment structure, we hoped to have them gain an overall sense of each utterance before proceeding to address the phonetic details that we thought might otherwise be vexing. This is a procedure that has been used in prior transcriptional coding in our laboratories, but its success in making coding more reliable is still uncertain—individual coders have given the procedure a mixed review, and systematic research on the effect of the procedure on reliability has not been conducted.

Additionally, the ordering of steps allowed for canonicity judgements (step 3) to be made prior to focus at the level of detail required for IPA transcription. The traditional IPA does not include canonicity symbols. Rather, use of the IPA *assumes* canonicity, and in the case of infant vocalizations this assumption is often inappropriate.

Supplementary sample—In general, the procedures utilized for the primary sample were also used in coding of the supplementary sample. However, the 6-step procedure used with the primary sample was abbreviated to 3 steps for the supplementary sample. The coders were instructed to:

1. record the syllable structure of each utterance by indicating all nuclei (Ns or vowel-like elements);
2. transcribe each utterance in LIPP™ using IPA (the number of transcribed vowels was required to correspond one-to-one with N codings from step one); and
3. make confidence judgements on the 1–7 scale for each IPA segment coded (i.e. how confident are you that the phonetic symbol you transcribed represents the sound you heard?).

Again, the coders were instructed to make use of all IPA symbols (including those from languages other than English) and diacritics when appropriate.

Calculations

The average proportion of canonical structures judged to occur within each of the 30 utterances of the primary sample was calculated across all 8 coders—this was the ratio of canonical Cs and Ns to all Cs and Ns at each utterance.³ Averages across transcribers for structure, canonicity, and transcription confidence judgements were also calculated for each of the 30 utterances. After an initial correlational analysis, it was decided that 2 of the 6 judgement levels would be omitted from the regression procedure. We chose to exclude structure judgements and structure confidence judgements, in part because step (5) required recoding in the greater detail afforded by IPA transcription of precisely the Cs and Ns coded in step (1). Thus, step (1) elements were, in essence, fully inferable from step (5) elements. Further, transcription confidence could be presumed to encompass the judgements made at the structure confidence level. This relation was reflected in high correlations observed between transcription and structure confidence judgements.

Alignment of codes—Due to differences in number of segments transcribed, choice of segments transcribed, and level of detail portrayed by coders, it was necessary to utilize a systematic alignment procedure in preparation for agreement analysis on each utterance for both the primary and the supplementary sample. Briefly, transcriptions were merged one by one in LIPP™. The second author's transcriptions were used as comparators and were assigned to the target row in LIPP™. Each of the 7 other coders' transcriptions were assigned to the transcription row.

As specified in Oller and Ramsdell (2006), 4 principles of alignment were followed. First, by the *nucleus alignment first principle*, vowel-like segments were aligned with other vowel-like segments first; then consonant-like segments were aligned with other consonant-like segments. The *strict-order principle* prevented re-ordering of segments transcribed by coders during alignment. The *matched segment principle* required that transcriptions with the same number of vowel-like and consonant-like segments ordered in the same way be aligned correspondingly, so that the vowel-like and consonant-like segments matched across codes. Finally, if there were different numbers of vowel-like and consonant-like segments, or if the vowel-like and consonant-like segments were ordered differently across transcriptions, the *minimum discrepancy principle* stated that the segments be aligned in such a way that those with maximally similar phonetic features be matched so as to produce minimum discrepancy between the aligned transcriptions. Table II shows alignment for one of the utterances. The transcriptions were aligned first across coders, and then the structure, canonicity, and confidence codings were aligned in the same pattern.

Weighted transcription agreement measure—The agreement measure to be considered here was calculated by a programme written in LIPP™ analysis language (LAL) and extensively described in Oller and Ramsdell (2006). The key principles of the analysis measure are summarized here.

³There was, in fact, wide variation in transcriber criteria for the assessment of canonicity. The mean utterance canonicity value for the 8 transcribers ranged from .90 to .25. This range could presumably have been limited by training to specific criteria of canonicity judgement based on work with many exemplars of infant utterances. However, it was our goal to assess intuitive responses both in terms of canonicity and confidence judgements. Hence, coders made their own decisions about how best to interpret the canonicity definition after it was provided along with a few example utterances during the training period. In the future we hope to conduct research to compare reliability for canonicity judgements in three circumstances: (a) as in the present work, with minimal criterion setting through training, (b) with much more rigorous training to limit the variation in criteria among coders, and (c) with purely instrumental acoustic canonicity judgements. Approach (c) will only be possible to implement after further specification of acoustic criteria for canonicity (see note 2).

The measure determines a proportion of agreement for each utterance, weighted segment-by-segment in terms of featurally specific degree of similarity between the transcriptions. The primary difference between this measure and the more commonly used percentage of agreement is that the latter is traditionally unweighted, treating all segment disagreements equally. Thus, in an unweighted approach disagreements on any segment result in a “0” score, while in the present procedure each disagreement in the transcription of a segment is weighted by the degree of the discrepancy in phonological features. Once agreement calculations were obtained using the weighted agreement procedure on the 7 coder comparisons for each utterance in the primary sample (or on the 3 coder comparisons for each utterance in the supplementary sample), they were averaged to determine the overall weighted transcription agreement for the group, which can be viewed as a general measure of transcription similarity between coders.

The *totally distributed weight principle* ensured that agreement between aligned segments was in every case weighted on a scale of 0–1, where 0 meant that no phonetic features were shared between the segments, and 1 meant all phonetic features were shared. The *equal steps principle* required division of the scale into equal intervals corresponding to each feature or sub-feature in accord with standard feature geometry assumptions (McCarthy, 1988). The result of application of these principles was that an ordinal scale based on phonological principles was treated as an interval scale of agreement from 0–1. The weighting procedure includes many additional provisions designed to incorporate and account for principles of phonological similarity and markedness (see Oller & Ramsdell, 2006).

The analysis focused on IPA base symbols plus only selected diacritics that were utilized systematically by the transcribers (those indicating nasalization of vowels, aspiration, and syllabification). Many additional diacritics were employed by the transcribers, including, but not limited to, symbols for tone, length, vocal quality, implosion, glottalization, retroflexion, and final consonant release. The analysis could be adjusted to account for disagreements on all diacritics of the IPA in the future. For our present purposes, however, it was deemed inappropriate to programme the many required adjustments to account for all IPA diacritics; transcribers were not encouraged to code every possible detail and showed great variability in their attention to diacritics other than those accounted for in the analysis.

To exemplify the measure, the general consonant features of place, manner, and voicing are each given a weighting of .333. Within each of these general features there are sub-features. For example, place disagreements come in 3 degrees, the largest of which ([p] to [k] for example) produces the full .333 reduction in agreement score for place, and the smallest of which ([k] to [c] for example) produces a reduction of only one-third as much (.0833). We recognize of course that non-linearities are surely involved in the scale since the impact of such discrepancies could hardly correspond in degree of perceived discrepancy precisely to the assigned score reductions. We reason, however, that the natural ordinal relation among featural disagreement types implemented through the equal steps principle should provide a workable approximation to an interval scale. And, as emphasized by Oller and Ramsdell (2006), user programmability of analysis within the LIPP™ software provides the basis for rapid update of the weighting scheme as new phonetic theory or perceptual research provides guidelines for an improved scaling of score reductions, with greater empirical grounding in data on perceptual similarity of phonetic elements.

The key criterion by which we gauge success of the approach is whether the weighted measure provides a measurable improvement over an unweighted measure. The results reported in Oller and Ramsdell (2006) confirm the improvement by showing that the weighted scheme yields agreement values that are well-distributed across the scale from 0–1, while an unweighted analysis of the same data yields very skewed distributions of values with dramatic floor effects.

The average agreement for the 30 utterances in the infant vocalization sample under the unweighted approach was .19, a value that is very low because more than half the utterances yielded 0 agreement—a floor effect. In contrast, using the weighted agreement analysis, the average overall transcription agreement was .60. The analysis procedure yielded a wide range of agreement across utterances transcribed. In the prior paper it is argued that the more evenly distributed values across the weighted agreement measure should provide a more stable basis for analysis of agreement (including correlation with other variables) than the unweighted measure with its skewed distribution.

The measure is reported below in terms of two components: global structural agreement and featural agreement. The two are combined in calculation of overall weighted transcription agreement. Global structural agreement measures whether both segment slots for any segment comparison (within columns of codes aligned across transcribers) contain segments. If a slot contains a segment for each transcriber, then a global structural agreement value of 1 is assigned. If only one of the transcriptions contains a segment, the pairing receives a value of 0 for the global structural agreement. Featural agreement measures featural similarity of segments in slots containing a transcribed segment by both coders, in accord with the principles described above. Overall weighted transcription agreement takes into account both global structural and featural levels of agreement by multiplying the two agreement values.

Analysis of the primary sample in two stages: Addressing the role of canonicity in transcription reliability

The utterances of the primary sample were carefully selected to represent a wide range of vocalization types occurring in this infant at the beginning of the canonical stage. As it turned out, 9 of the utterances were transcribed with no supraglottally articulated Cs. As a result, there was no possibility in these particular utterances for occurrence of non-canonical C to N transitions. The speculations reviewed above about the effect of canonicity on transcription agreement are focused almost entirely on the impact of non-canonical transitions.

Consequently, we reasoned that the 21 utterances whose transcriptions included supraglottally articulated Cs should yield a more useful test of the effect of canonicity on transcription agreement than the whole 30-utterance sample, which might show lesser effects due to the fact that it included utterances that had no possibility of showing differentiation on the presumed key factor, namely C to N transition canonicity.

This reasoning led us to conduct two analyses. One was conducted on the 30-utterance sample (with nominally 819 segment-level data points for coding comparisons at each level of utterance and confidence judgements), and a second analysis (which we presume to be the more enlightening one) on the sub-set of 21 utterances (with nominally 700 data points for each comparison) that included supraglottally articulated Cs and consequently all of these utterances included the possibility of variation in C to N transition canonicity.

Results: 30-utterance analysis for the primary sample

Ordinary least squares multiple regression was used to determine the influence of utterance canonicity and coder confidence on transcription agreement for the primary vocalization sample. The weighted transcription agreement values provided the dependent variable and utterance canonicity (as determined by the proportion of canonical structures present in each utterance, averaged over the coders), coder canonicity confidence, and coder transcription confidence provided the independent variables. The 819 (117 segment slots times 7 coder pairings) comparisons at each level were averaged across each utterance and across coder pairings, yielding 30 utterance-level data points for each 0-order correlation. The means, standard deviations, and correlations among all the variables are given in Table III.

Means and standard deviations for the scores on each measure are provided at the bottom of Table III. The means indicate that the average agreement across the 30 utterances was .60; thus 3/5 of phonological features were shared on average across transcriptions. The utterance canonicity mean suggests that nearly half the segments in the sample were deemed, on average, to be canonical. The confidence judgement measures, based on the 1–7 rating scale, suggest that observers had slightly higher confidence in canonicity judgements than in transcription judgements.

Goal 1: Was utterance canonicity significantly related to transcription agreement?

The zero-order correlation between utterance canonicity and transcription agreement was significant ($r=.376$, $n=30$, $p<.05$, 2-tailed), indicating that, as has been presumed in much prior writing, canonicity is a factor in transcription agreement. When utterances were non-canonical, they tended to produce low transcription reliability.

The effect of canonicity on transcription agreement can be illustrated with regard to Figure 1. For the 8 coders, the transcriptions corresponding to the canonical syllable (including the initial consonant-like element and the nucleus but not including the subsequent consonant-like element) were: pa, bɑ, ba, ba, bæ, pɑ, ba, ba and the transcriptions corresponding to the non-canonical syllable (also from the initial consonant or consonants through the nucleus) were: ba, bwæ, bwɑ, blɑ, βe, pwa, mwæ, mwæ. One possible interpretation of the result on these syllables is that the non-canonical syllable's slow transition apparently produced disagreement about whether there was a glide or other consonant following the initial consonant-like element. Further, the initial consonant in the non-canonical syllable was heard as a fricative or a nasal in three cases, whereas it was always treated as a stop in the canonical syllable.

Goal 2: Was transcription confidence significantly related to utterance canonicity?

The zero-order correlation between transcription confidence and utterance canonicity was also significant ($r=.671$, $n=30$, $p<.001$, 2-tailed). The result is consistent with reports of laboratory staff, who have indicated often in our prior research that they feel low confidence in transcriptions of utterances that include non-canonical syllables.

Goal 3: What were the relative contributions of utterance canonicity and confidence judgements in predicting transcription agreement?

As indicated in Table III, in addition to utterance canonicity, canonicity confidence and transcription confidence were also significantly related to transcription agreement. The order of the relationships between the independent variables and transcription agreement, beginning with the strongest, was transcription confidence ($r=.566$, $n=30$, $p<.01$, 2-tailed), followed by canonicity confidence ($r=.527$, $n=30$, $p<.01$, 2-tailed), and utterance canonicity ($r=.376$, $n=30$, $p<.05$, 2-tailed). As confidence increased, transcription agreement also increased. As presence of non-canonical structures in an utterance increased, transcription agreement decreased.

Despite the high correlations among these variables, examination of the regression results indicated that there was no extreme multicollinearity in the data (all variance inflation factors were less than 2.5) nor were there any influential (outlier) data points. The lack of multicollinearity indicates that the 3 variables (canonicity, canonicity confidence, and transcription confidence) were not so highly correlated as to suggest that they were acting as one in predicting transcription agreement, and it was thus justified to compare the individual regression coefficients.

In order to compare the contributions of the independent variables to transcription agreement, they were entered into a regression equation with the overall weighted transcription agreement value as the dependent variable. A block-entry (or hierarchical) approach was taken in the

estimation of the equation with utterance canonicity entered alone in Model 1. The two confidence variables, canonicity confidence and transcription confidence, were then added to the equation in Model 2. Utterance canonicity was entered into the regression equation first, and independently, because the experimental design required coders to judge utterance canonicity first and also because we had originally hypothesized that utterance canonicity might be the driving force in transcription agreement. Table IV indicates the results of the analyses for each model.

The regression results indicated that both utterance canonicity, explaining 14.1% of the variance [$F(1, 28)=4.614, p<.05$], and coder confidence, explaining an additional 24.3% of the variance [$F(3, 26)=5.421, p<.01$], independently explained a statistically significant amount of variance in weighted transcription agreement, a total of 38.5%. Not only did the confidence variables explain additional variance when added to the equation in Model 2, but they also accounted for more than 1.5-times as much variance in transcription agreement as did utterance canonicity alone in Model 1. When entered independently into the regression equation, utterance canonicity had significant unique influence on transcription agreement, with $\beta=.376$. When both canonicity and confidence variables were present in the model, the set of variables explained a statistically significant proportion of variance in transcription agreement, but none of the independent variables exhibited a significant unique relationship to transcription agreement (i.e. none of the coefficients was statistically significant). Nevertheless, the magnitudes of the standardized coefficients (.306 and .391 for utterance canonicity and transcription confidence, respectively) hint at a substantive effect of confidence on transcription.

Results: 21-utterance analysis for the primary sample

As explained above, 9 of the 30 utterances in the primary sample had no supraglottally articulated Cs, according to the transcriptions, and were consequently irrelevant to assessing the role of canonicity of articulatory transitions as a determiner of transcription agreement. The 21-utterance analysis excluded these 9 utterances, reducing the total comparison from 819 to 700 (100 segment slots in the utterances times 7 coder pairings). As expected the results showed a markedly stronger relation in the 21-utterance analysis between utterance canonicity and transcription agreement than in the 30-utterance analysis.

In the 21-utterance analysis, the canonicity variable consisted of the average proportion of canonical transitions within each utterance (that is, canonical nuclei occurring without adjacent supraglottally articulated Cs were not included in sums to determine the average proportion), while the confidence variables were computed the same way as in the 30-utterance analysis. The means, standard deviations, and correlations among all the variables are given in Table V. The means at the bottom of the table suggest that transcription agreement for the 21 utterances was slightly lower than for the 30 utterances and that the 21-utterance sample had slightly lower average canonicity per segment than the 30-utterance sample. Mean confidence values changed little across the analyses.

Goal 1: Was utterance canonicity significantly related to transcription agreement?

The zero-order correlation between utterance canonicity and transcription agreement was significant ($r=.626, n=30, p=.002, 2$ -tailed), indicating more strongly than in the 30-utterance analysis, the role of canonicity in transcription agreement. When utterances were non-canonical as a result of having non-canonical transitions, they tended to produce low transcription reliability.

Goal 2: Was transcription confidence significantly related to utterance canonicity?

The zero-order correlation between transcription confidence and utterance canonicity was also significant ($r=.645$, $n=30$, $p=.002$, 2-tailed). The result is similar to that of the 30-utterance analysis and is again consistent with reports of the limiting effects of non-canonical syllables on transcription agreement made by laboratory staff.

Goal 3: What were the relative contributions of utterance canonicity and confidence judgements in predicting transcription agreement?

While the relation between transcription confidence and transcription reliability changed little from the 30-utterance analysis to the 21-utterance analysis, the large shift in the zero-order correlation between utterance canonicity and transcription agreement provided a basis for a notable change in the regression results. All the independent variables were statistically significantly correlated with transcription agreement, such that as utterance canonicity increased (fewer non-canonical transitions), transcription agreement increased, and as confidence increased (both canonicity and transcription confidence), transcription agreement increased. While the independent variables showed significant correlations with one another, examination of the regression results indicated that there was no extreme multicollinearity in the data (all variance inflation factors were less than 2.5), nor were there any potentially influential data points, both noteworthy facts with 21 utterances represented in the correlations.

The regression results (shown in Table VI) indicated that both canonicity, explaining 39.2% of the variance [$F(1, 19)=12.243$, $p<.01$] and coder confidence, explaining an additional 23% of the variance [$F(3, 17)=9.327$, $p<.01$] independently explained a statistically significant amount of variance in weighted transcription agreement, a total of 62.2%. Here, however, utterance canonicity in Model 1 explained nearly twice as much variance as the confidence variables when entered in Model 2, a pattern contrasting sharply with that of the analysis with 30 utterances. Additionally, unlike in the 30-utterance analysis, utterance canonicity showed a significant *unique* effect on transcription agreement in the 21-utterance analysis, when entered into the full regression equation with the coder confidence variables ($\beta=.380$). Furthermore, Table VI shows that canonicity confidence narrowly missed significance ($t=2.05$, $p=.056$), but, again, the magnitude of its standardized coefficient (.478) suggested a substantive effect on transcription agreement.

Transcription confidence (comparing Tables IV and VI) dropped in predictive power dramatically from the 30- to the 21-utterance analysis, being substantially supplanted by canonicity as a predictor of transcription agreement. However, the zero-order correlation between canonicity confidence and transcription confidence (Table V) suggests that the two were closely related. Thus, the 21-utterance analysis produced results conforming much more closely to our expectation about a role for canonicity in transcription agreement than the 30-utterance analysis. These results suggest that earlier speculations about the critical role of aberrant articulatory transitions in reducing transcription agreement may be correct.

Additional results

Goal 4: Assessing the role of canonicity, non-nativeness, and vocal quality aberrations in transcription agreement for infant vocalizations in the primary sample

Results of specific t-test comparisons are depicted in Figure 2. Figure 2(a) shows the comparison between agreement values for transcription of non-canonical utterances vs canonical utterances. This analysis was conducted at the utterance level because transition aberrations cannot be localized to a single segment in terms of perceptual effects, but tend instead to affect whole syllables and often affect the number of perceived syllables in an utterance. In addition, the analysis was limited to the 21 utterances that included at least one

transcribed supraglottally articulated C (7 non-canonical, 14 canonical utterances). Canonical utterances in this case were determined in a review by the authors of the 21 utterances, focusing on effects of perceived canonicity at the utterance level.⁴ Therefore, the total number of comparisons was ~700 across all coder pairings.

The effect of canonicity was robust. Non-canonical utterances showed statistically significantly lower agreement than canonical utterances: $t(6)=7.584$, $p<.01$. The effect size in Cohen's d was very large ($d=2.867$). The average overall agreement on canonical utterances was .62 compared to .45 on non-canonical utterances.⁵ All 7 coder pairings showed higher agreement on canonical utterances, and there was no overlap in the distributions—that is, the highest overall transcription agreement for any pairing on non-canonical utterances was lower than the lowest overall transcription agreement for any pairing on canonical utterances.

The role of non-nativeness of sounds transcribed with respect to the primary language of the transcribers (American English) is portrayed in Figure 2(b). We defined non-native sounds to include any element *not* in the phonetic (including allophonic) repertoire of American English. The analysis was keyed to the standard coder's judgements of non-nativeness since the analysis format, as in all the other analyses in the study, compared each of the other coders to the standard.⁶ The following IPA base symbols meeting the definition of non-native actually occurred in the transcriptions of the standard coder: ɪ, ʏ, ʉ, ʊ, ɣ, β, x, ɸ, R, ʁ, Φ, Gɸ, q. This analysis was conducted through an adaptation of the LIPP™ weighted reliability programme operating at the segment level, with averages across the entire sample rather than at the utterance level. There were 24 non-native segment tokens and 93 American English tokens in the standard coder's transcriptions. The total number of comparisons was 819 across all coder pairings.

Non-native sounds in transcriptions resulted in statistically significantly lower agreement than transcription of sounds from the American English repertoire: $t(6)=-7.142$, $p<.01$. And again, there was a very large effect size for this comparison, Cohen's $d=3.955$. Non-nativeness of transcribed sounds was not confounded with the distinction between canonical and non-canonical utterances; the number of non-native transcribed segments was ~1.1 per utterance, for both canonical and non-canonical utterances.

The analysis of the effect of vocal quality aberrations on transcription agreement was conducted at the utterance level because aberrant vocal quality was a property that extended across large stretches of utterances rather than being confined to individual segments. Of the 30 utterances coded, 10 demonstrated aberrant vocal quality and 20 demonstrated typical vocal quality. Results are shown in Figure 2(c). The authors reviewed the utterances and made the determination of aberrant vocal quality based on perceived significant occurrence within each

⁴The utterance canonicity judgements for this t-test analysis were not identical to the ones utilized in the correlational analyses. For the t-test analysis we sought to indicate lack of canonicity in terms of articulatory transitions that auditorily seemed particularly disruptive to the rhythmic structure of whole utterances. In contrast the judgements for the correlational analyses were made at the level of the segment with no particular attention to the utterance as a whole.

⁵In a separate descriptive analysis utilizing the segment-by-segment judgements from the correlational analyses, we split the distribution by utterances with canonicity values above the mean and those with values below the mean. Results were similar to those reported in the main text for utterances categorized as canonical or non-canonical—in the split plot analysis, utterances with canonicity values above the mean had a similar and slightly larger advantage in transcription agreement over utterances with canonicity values below the mean (.65 vs .45, respectively).

⁶The 7 comparator coders utilized a variety of phonetic symbols corresponding to sounds not occurring in American English in addition to those listed for the standard coder. Among the standard coder's non-native symbols were many that occurred multiple times across the comparator coders. The additional non-native symbols not utilized by the standard coder were used very infrequently across the other coders, the great bulk of them exactly once. To simplify the LIPP™ analysis program (which would have had to be elaborated greatly to incorporate all the infrequently occurring symbols as non-natives), we resolved to key the analysis on the standard coder's utilization of non-native symbols.

utterance of falsetto, breathiness, vocal tremor, or pressed voice. The total number of comparisons was (as with the non-nativeness analysis) 819.

Transcription of utterances with aberrant vocal quality (many of which were squeals, typically showing falsetto register as judged by the coders) also resulted in statistically significantly lower agreement than utterances with typical vocal quality: $t(6)=-9.508$, $p<.01$. Again there was a very large effect size (Cohen's $d=3.529$) for the analysis. There was, however, a tendency for aberrant vocal quality to occur more often in the non-canonical than the canonical utterances (by a factor of 4). Consequently, the possibility is raised that effects seen for canonicity on transcription agreement in the primary sample could have been caused by vocal quality variations, at least in part. The independent effect of canonicity is evaluated further in the supplementary sample below, where vocal quality aberrations were excluded.

Goal 5: Item analysis for the primary sample

To provide additional perspective on the primary sample regarding breakdown in transcription agreement, an item analysis was conducted. Based on instructions given in the coding protocol, there was substantial usage across the transcribers of IPA symbols pertaining to phonetic elements of languages other than English. Twenty-one per cent of transcribed segments corresponded to base symbols not used in phonetic description of English. Diacritics were also utilized, especially the nasalization diacritic for vowels, the non-syllabification diacritic to indicate semi-vowel status, as well as indicators for aspiration, retroflexion, length, tone, and vocal quality. As indicated above, however, the LIPP™ analysis did not account for all diacritic usage in calculating transcription agreement, because there was substantial variation in the tendency of the transcribers to incorporate judgements on such factors as tone and length. Examples of common V and C disagreements on base symbols found in the paired transcriptions are shown in Table VII, ranked by their frequency of occurrence.

A final analysis on the primary sample compared transcription agreement for consonants and vowels. As in the other comparisons, all aligned slots were included, even if there was no segment indicated for one of the transcribers. Thus, the data on Cs and Vs pertain to overall transcription agreement (structural agreement times featural agreement). The results are averaged across the entire sample of segments rather than across utterances. There was an average of 62 V slots and 55 C slots per coder pairing. The total number of comparisons of paired segment slots was 819.

The item analysis revealed that transcribed Cs resulted in statistically significantly lower average agreement across coders ($M=.504$, $SD=.071$) than transcribed Vs ($M=.644$, $SD=.044$): $t(6)=3.692$, $p<.01$. The stability of the pattern is demonstrated by the large effect size ($d=.865$). It is worth noting that the greater disagreement on Cs was largely due to a greater tendency for Cs than Vs to be in slots where only one of the two transcribers indicated a segment was present (i.e. the other transcriber indicated the presence of no segment at all). This tendency was reflected in significantly higher vowel than consonant agreement on the global structural measure (vowel structure agreement=.81, consonant structure agreement=.66, $t(6)=-4.051$, $p<.01$, $d=.464$). For featural agreement only (where both transcribers indicated the presence of segments), Cs and Vs had similar levels of agreement (vowel featural agreement=.79, consonant featural agreement=.76, $t(6)=-1.446$, $p=.198$).

Goal 6: Analysis of the supplementary sample to assess generalizability of the effects of canonicity and coder confidence on transcription agreement

Figure 3 compares the overall weighted agreement (structural times featural agreement) on canonical vs non-canonical vocalizations for the supplementary sample across the three coder comparisons. The data for the supplementary sample offer a more targeted comparison than

the primary sample for infant utterances with canonical and non-canonical articulatory transitions. Utterances from the supplementary sample were pre-selected on the basis of their being sharply distinct with regard to presence of canonical or non-canonical transitions. Further, they offer the opportunity to assess the generalizability of the effects of canonicity and coder confidence on transcription agreement because they involve additional transcribers and utterances from 3 additional infants. Finally, the supplementary sample was controlled to eliminate possible confounds between vocal quality and canonicity or between non-nativeness and canonicity. As such, the predicted effect of canonicity on transcription agreement could be interpreted as being attributable to canonicity with more confidence.

As with the t-test comparison described above with respect to Figure 2(a), this analysis was conducted at the utterance level because transition aberrations related to non-canonicity tend to affect transcription of whole syllables, often affecting both transcribed segments and the number of perceived segments or syllables in utterances. Data averaged over the 3 transcriber pairings included 384 comparison points (an average of 128 segment slots in the utterances times three coder pairings).

With the new supplementary sample and the new coder comparisons, the effect of canonicity was robust. Non-canonical utterances showed statistically significantly lower agreement ($M=.559$, $SD=.108$) than canonical utterances ($M=.763$, $SD=.087$) based on averaged agreement values for the 3 coder comparisons: $t(18)=4.64$, $p<.001$. The effect size in Cohen's d was very large ($d=2.068$). All 3 coder pairings showed higher agreement on canonical than non-canonical utterances at $p<.01$. In addition, as a further check on robustness of the effect of canonicity, we assessed the 2 new transcribers with respect to each other (aligning one of the coders as comparator and one as standard), and found that the canonical utterances yielded (as in all prior such comparisons made for this research) statistically reliably higher ($p<.01$) transcription agreement than the non-canonicals.

The supplementary analysis also showed that coder confidence was higher for canonical ($M=5.333$, $SD=.751$) than non-canonical utterances ($M=4.570$, $SD=.337$). This result was statistically significant: $t(18)=-2.93$, $p<.01$ and the effect size in Cohen's d was large ($d=1.311$), further confirming the reliability of the relation of coder confidence and transcription reliability.

Discussion

Canonicity as a predictor of transcription reliability

The most important goal of the present research was to assess the predictive value of canonicity in transcription agreement. This factor has not previously been documented in the literature as influencing transcription agreement, even though it has been hypothesized that non-canonical utterances are particularly hard to transcribe since the 1970s when the notion of the canonical syllable was formalized (Oller, 1978).

Several kinds of analyses from the present work bear on the goal. Two correlational analyses were conducted, one on the entire primary sample of 30 utterances and another on 21 of those utterances. As we hypothesized, in both analyses there were significant positive correlations between canonicity and transcription agreement. Also in both cases, canonicity accounted for significant R-square in the Model 1 regression analyses; proportions of variance that were accounted for by utterance canonicity in the 30-utterance and 21-utterance analyses were .14 and .39, respectively. These results confirm our contention that canonicity is a factor that may play an important role in transcription agreement.

In addition, the fact that the correlation between utterance canonicity and transcription agreement was higher in the 21-utterance analysis than in the 30-utterance analysis (specifically 2.5-times as much variance in transcriber agreement) supports the speculation that the primary factor in canonicity that hampers transcription concerns violations of the rapid articulatory transition requirement of canonical syllables. In the 30-utterance sample, there were 9 utterances that had no supraglottally articulated Cs and, consequently, because they could not have C-to-V transitions, there was no differentiation among them on the rapid articulatory transition requirement of canonical syllables. It would appear that these 9 utterances, when included in the correlations (as they were in the 30-utterance analysis), tended to weaken the apparent effect of canonicity. Thus, we deem the 21-utterance analysis to be the more revealing as a test of the hypothesis that canonicity may affect transcription agreement.

Further, the effect of canonicity was assessed at the utterance level with a t-test across the 7 coder pairings, again averaging over the 21-utterances. The averaged data on all coder pairings showed a significant and large effect where canonical utterances had higher agreement than non-canonical utterances.

Finally, the robust effect of canonicity on transcription agreement was also supported by results from the supplementary sample conducted with a new set of vocalizations from 3 infants as transcribed by additional coders. As with the primary sample, utterances judged to be canonical in the supplementary sample showed statistically higher transcription agreement than utterances judged to be non-canonical. In addition, the supplementary sample analysis helped to eliminate the possible interpretation that the effect of canonicity on transcription agreement might be due to a variable (vocal quality aberrations) partially confounded with canonicity in the naturalistically selected utterances of the primary sample. The supplementary sample utterances were selected to exclude vocal quality aberrations. As such, the overall results on canonicity provide strong evidence that low canonicity is a factor in low transcription agreement for infant vocalizations.

It should be emphasized that all analyses on the effect of canonicity were conducted at the utterance level, even though the effects of canonicity may in fact be localized to the syllable level. Individual utterances with multiple syllables sometimes had a mix of canonical and non-canonical syllables, particularly in the 21-utterances from the primary sample. Consequently, the size of the effects we found for canonicity on transcription agreement can be assumed to provide a lower limit estimate because the analysis did not take advantage of focus at the syllable level. With the supplementary sample also, only about half the syllables in the non-canonical utterances (17/35) included non-canonical formant transitions. Consequently, a syllable-level analysis, targeting only non-canonical and canonical syllables differentiated by the articulatory transition criterion in the sample, might show an even more powerful effect of canonicity on transcription agreement. We are currently working towards an update in the LIPP™ analysis language (LAL) to develop a capability that should in the near future facilitate syllable-level analysis.

In addition, the magnitude of the effects of canonicity in the present analyses on the primary sample may have been affected by variability in the judgement of canonicity itself. Coders varied greatly in how often they indicated segments to be non-canonical. The level of agreement among coders on canonicity itself might be substantially improved by more extensive and directive training, which were not a focus of the present study. Our approach here for the primary sample was to define canonicity and illustrate the definition during training sessions conducted within a single week. Given the variability among coders in their choice of a criterion of judgement for canonicity, it is reasonable to speculate that the correlations we obtained between canonicity and transcription agreement were lower than they might have been if the canonicity variable had been controlled through additional training. Further, a fully

instrumental acoustic approach to determining canonicity could in the future conceivably improve the judgement of canonicity and lay the basis for more precise correlational assessment of the role of canonicity in transcription agreement.

Variability in judgement of canonicity was artificially controlled in the supplementary sample. The second author categorized all the utterances that were to be included in the sample as canonical or non-canonical—these categorizations were confirmed subsequently by the first author. The categorization was auditory. It remains possible that acoustic analysis could improve the accuracy of canonicity judgements (once all the necessary acoustic definitional criteria for canonicity have been stabilized), and could consequently provide the basis for more accurate assessment of the role of canonicity in transcription agreement.

Transcription confidence as a predictor of canonicity

The second important goal of the present research was to assess and quantify the role of transcription confidence on perceived segments within utterances as a predictor of the perceived canonicity of the utterances. Again, this relation has been assumed to be important since the formulation of the idea of canonicity in the 1970s. In fact, the idea of canonicity was partly formulated as a result of reports from transcribers who felt low confidence in the transcription of certain utterances. Yet until now there has been no quantitative evidence that transcription confidence relates systematically to canonicity.

The results here show that the relation is reliable. In both 30- and 21-utterance analyses, the correlations between confidence and canonicity were statistically significant and the proportions of variance accounted for were .45 and .27, respectively. The results suggest that indeed, transcription confidence does predict canonicity systematically. As with the relation between canonicity and transcription agreement, if we could obtain a more reliable measure of canonicity, we might well expect an even higher correlation between canonicity and transcription confidence. Further, the results on the supplementary sample provided additional support for a solid relation between coder confidence and transcription agreement. Coders showed statistically reliably higher confidence on canonical than non-canonical utterances.

Regression analysis on canonicity and confidence variables as predictors of transcription agreement

The regression analyses presented in Tables IV and VI tested the possibility that the confidence variables accounted for additional unique variance in transcription agreement beyond that accounted for by canonicity. The two analyses also afforded the opportunity to evaluate relative contributions of canonicity and the confidence variables modelled two ways. Indeed both analyses showed increases in R-square accounted for when the confidence variables were added to the canonicity variable in Model 2. Although both confidence measures (canonicity confidence and transcription confidence) showed some-what higher zero-order correlations than canonicity alone in the 21-utterance analysis, the role of canonicity within the regression equation was stronger in the 21-utterance analysis, where canonicity played a statistically significant independent role in influencing transcription agreement. While the 30-utterance analysis showed an R-square change when confidence variables were added to the model that was more than 1.5-times the R-square accounted for by canonicity alone, in the 21-utterance analysis, the R-square change, while still significant, was not nearly as dramatic. Notably the total amount of variance accounted for by the full regression equation was considerably higher (.622) in the 21-utterance analysis than in the 30-utterance analysis (.385). The absolute increase in R-square change accounted for by the confidence variables was comparable in the two analyses (.243 and .230, respectively).

These facts provide additional reasons to suggest that the 21-utterance analysis may have provided a better focus on the aspect of canonicity that predicts transcription agreement than the 30-utterance analysis. It suggests that indeed, the primary problem with transcription reliability owing to canonicity may be related directly to articulatory transition aberrations. Transitions judged to be non-canonical in the sample were often perceived as slower than transitions typically found in mature speech, and other perceived non-canonical syllables possessed quasivocalic nuclei with so little vocal aperture as to prevent the occurrence of any substantial articulatory transition from adjacent consonant-like elements. The results from the supplementary sample provide additional support for the idea that articulatory transition aberrations may play an especially potent role in hampering transcription agreement. The results confirm the effects found with the 21-utterance analysis of the primary sample, showing that canonicity predicts agreement.

Reflecting on the results regarding the confidence variables, we suggest that transcribers are aware of a variety of factors, of which canonicity is one. Other factors entering into transcription confidence judgements could include awareness of sounds that are unfamiliar to the coders, awareness of possible effects of aberrant vocal quality, etc. Under Goal 4 (see next section) it was indicated indeed that both familiarity (or its inverse, non-nativeness) and vocal quality play substantial roles in transcription agreement. These factors have both been hypothesized as playing roles in transcription reliability based on prior research (Kearns & Simmons, 1988; Coussé et al., 2004). It makes sense that coder transcription confidence could encompass such factors and, thus, could correlate with transcription agreement.

The role of non-nativeness and vocal quality aberrations in transcription agreement: Familiarity as a potentially general factor in transcription agreement

Prior literature indicates that it is difficult to transcribe reliably sounds that are unfamiliar to transcribers or that are not present in transcribers' ambient language environment (Louko & Edwards, 2001; Coussé et al., 2004). Results of the non-nativeness analysis supported this idea because infant vocalizations with non-native-like sounds resulted in greater disagreement between transcriptions than those with sounds that exist within the phonetic repertoire of the transcribers' primary ambient language, American English. Non-native sounds transcribed and compared across coders showed an agreement of .44, as compared to .64 for the standard American English sounds transcribed and compared across coders. The result was both highly significant and corresponded to a very large effect size.

The non-nativeness result is consistent with the results of prior studies, and we are inclined to interpret the result as being related to the canonicity effect. Non-nativeness is a variable pertaining to familiarity of phonetic templates presumably developed through experience with particular languages. In invoking the idea of perceptual templates, we take account of a vast literature on perception indicating that degree of experience with auditory stimuli produces enhancement and stabilization of templates that are utilized in recognition of sounds (Werker & Tees, 1984; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Goldstone, 1998; Best & McRoberts, 2004). Just as nativeness of sounds can be viewed as pertaining to high experience with a particular type of perceptual event, and consequently a good match to stable templates of the listener, so canonicity can be viewed as pertaining to familiarity of phonetic templates. Speakers of a language accumulate considerable experience with canonical syllables. Non-canonical syllables, by contrast, can be characterized as not fitting well with the templates of listener experience. Thus, in both cases (non-nativeness and canonicity) it may be that transcription agreement is driven by the relative similarity of sounds that listeners attempt to interpret phonetically with respect to phonetic templates resulting from their personal listening experience.

Variations in vocal quality have also been reported to result in limited agreement across transcribers (Duckworth et al., 1990). Since infants produce a wide range of vocal qualities (from squeals and growls to yells and whispers), greater difficulties in transcription may result. The present data indicated that infant vocalizations with aberrant vocal quality did result in significantly greater disagreement across transcriptions. These vocalizations showed an average overall weighted transcription agreement of .50, while those with primarily modal vocal quality had an average of .65. The effect size once again was very large.

We envision at least two mechanisms by which vocal quality aberration might reduce transcription reliability. The first concerns distortions accompanying various sorts of aberrant vocal qualities: for example, in falsetto voice, harmonics are so widely spaced that formant values may be difficult to discern (Robb & Carcase, 1995; Kent & Read, 2002). The second mechanism invokes again the notions of familiarity and template matching. Speech is normally produced in modal phonation, and consequently listeners have much more experience with syllables presented in modal phonation than in any other register. Consequently, it may be the case that aberrant vocal quality corresponds to vocal sequences that are inherently poorly matched to the relevant perceptual templates, learned and stabilized in experience, and utilized in identification.

Our reasoning here suggests that lack of similarity of vocalizations that are to be transcribed with templates corresponding to listening experience may reduce transcription agreement in three related ways—through low canonicity, through non-nativeness, and through aberrations in vocal quality. It is also of interest that when transcription confidence judgements are made, they may take all these factors into account. Transcribers in the study indeed indicated that canonicity, nativeness, and vocal quality were all auditorily noticeable.

A note of caution about the results on vocal quality and non-nativeness from the present study should be raised, however. In the naturalistically selected utterances of the primary sample, the vocal quality factor was partially confounded with the canonicity factor—more than half the utterances designated as non-canonical had vocal quality aberrations, while only 14% of canonicals showed such aberrations. In analysis of the supplementary sample, we controlled for vocal quality aberrations by excluding them from the sample. Analysis of the supplementary sample helped substantiate a role for canonicity in transcription agreement. However, the vocal quality effect in the present study can be interpreted as partially confounded with canonicity.

The confound for the effect vocal quality runs deeper yet, because transcribed non-native sounds occurred more often in utterances with vocal quality aberrations than in utterances without such aberrations (1.1 per utterance to .65 per utterance). Consequently, neither non-nativeness nor vocal quality aberrations can be unambiguously interpreted as causing low transcription agreement in the present study, because they may have operated in concert.

Canonicity as a factor in transcription agreement can be more confidently asserted based on the present work, first because there was no confound with non-nativeness in the primary sample, and secondly because the supplementary sample explicitly controlled both vocal quality and non-nativeness. Utterances with aberrant vocal quality were excluded and utterances expected to yield transcriptions including non-native sounds were intentionally concentrated among the canonical utterances, where their effect would presumably have been to reduce the difference favouring agreement in canonical as opposed to non-canonical utterances.

On vowel and consonant transcription

Prior researchers have indicated that vowels may be more difficult to transcribe reliably than consonants (Pollock & Berni, 2001; Stoel-Gammon, 2001). Our study found, however, that

vowel-like segments yielded higher agreement (.64) across transcriptions on infant vocalizations than consonant-like segments (.50). The difference between our results and those reported by the cited prior investigators could be due to a variety of factors, with three that seem particularly likely. First, our transcriptions were based on vocalizations from an infant in the first year of life, while the prior reports refer primarily to transcription of meaningful speech in older children. Secondly, the calculation of the overall agreement values in our research combined two cases, one where 2 transcribers may or may not have agreed about whether segments existed at all at a particular alignment slot, and another where both transcribers indicated a C or V segment for the slot, while the prior studies appeared to be primarily focused on cases where both transcribers indicated a segment. Thus, our results appear to have suggested a higher agreement rate on vowels because vowels were more often included in both compared transcriptions at a slot than consonants were. And, thirdly, our comparisons were based on a weighted reliability measure that was not previously available, and, therefore, resulted in agreement values that were scaled differently than in prior work. The present research provides a first glance at the relative agreement levels for transcription of consonant-like and vowel-like sounds in infant vocalizations as seen through an overall weighted measure.

Potential advantages of the weighted transcription agreement measure

The transcription agreement measure utilized here (Oller & Ramsdell, 2006) weights each disagreement between two transcriptions by degree of discrepancy in phonological features, and is specifically designed to incorporate and account for principles of phonological similarity and markedness. A primary reason for introducing the new measure was that we presumed it might provide a more stable basis for correlation than the traditional unweighted approach. The research provides empirical support for the stronger scaling properties of the weighted measure in analysis of transcription comparisons than unweighted measures (Oller & Ramsdell, 2006).

The 21-utterance analysis seemed to provide clearer support than the 30-utterance analysis for the suggestion that the weighted transcription agreement would provide greater predictive power in correlational analysis. Exploration of both samples yielded small differences between regression analyses for weighted vs unweighted agreement values. For the weighted agreement values under the 30-utterance analysis, the R-square for the full regression equation was .385, while the unweighted R-square was a bit lower, .350. For the weighted agreement values under the 21-utterance analysis, the R-square for the full regression equation was .622, while the unweighted R-square was also only a bit lower, .579. This tells us that, under both analyses, there was only a small increase in R-square with the weighted measure over the unweighted measure.

The more notable difference between the outcomes based on examination of weighted and unweighted methods was not in the R-square change overall, however, but in the R-square accounted for by canonicity: For the weighted measure in the 21-utterance sample full model, the R-square was .392, while the unweighted measure yielded only .167. Thus, with the weighted measure, canonicity accounted for more than twice as much variance as it did with the unweighted measure of transcription agreement. This difference was also reflected in greater *relative* R-square accounted for by canonicity than by confidence under the weighted measure. Here, the weighted measure in the 21-utterance analysis yielded a total R-square of .622 in the full model, and canonicity (with R-square of .392) accounted for 63% of that total. Using the unweighted transcription agreement measure, the total R-square was .579 in the full model and canonicity (with R-square of .167) accounted for only 29% of the total.

This pattern of relative outcomes for canonicity and confidence suggests that the weighted measure may have provided a special window into the role of canonicity in transcription

agreement. We are inclined to suggest that the value of the weighted approach could be primarily in its ability to *differentiate* factors that account for transcription agreement, although the results also hint at a possibly overall greater power of the weighted measure than the unweighted one in supporting correlations between transcription reliability and other variables.

Differences between canonicity and confidence level judgements

Our research approached transcription agreement from two related angles, both of which were accessed through judgements made by the transcriber-subjects. Canonicity judgements of transcribers are conceptually different from confidence judgements in that the former are *assessments of utterances*, while the latter are *assessments of internal states of the transcribers*. Further, in the case of canonicity judgements (labelled *utterance canonicity* in Tables III–VI), the transcribers were instructed to focus assessment on a particular set of features of the stimuli (the features that pertain to canonicity). In the case canonicity confidence, coders were also asked again to focus attention on the characteristics of sounds that pertained to canonicity. In the case of transcription confidence, however, there was no instruction to limit the features of the stimuli that could influence the confidence judgements—many features including canonicity, non-nativeness, vocal quality, signal-to-noise ratio, etc., could potentially factor into the judgement.

Consider, for example, the complexity of utterances (e.g. number of alternations of place or manner of articulation within each utterance). Coders were encouraged to allow any feature of utterances that they thought may have affected their transcription to enter into their transcription confidence judgements, and so complexity of utterances was among the factors that could have been expected to play a role in confidence judgements. In contrast, transcribers were instructed not to allow such changes in complexity of articulation to affect *canonicity judgements*; instead they were asked to focus on the well-formedness of each segment or syllable at the point of canonicity judgement. Similarly, according to the coding protocol, familiarity with language-specific articulations (i.e. the degree to which particular articulations such as trills or front rounded vowels might pertain to languages known by the transcriber) occurring in stimuli was a legitimate basis for making transcription confidence judgements, but not canonicity judgements. To limit the role of familiarity with sounds of particular languages in canonicity judgements, coders were asked instead to focus on infrastructural features that pertain to all natural languages and apply to all types of syllables and segments that occur regularly in any language. In accord with this instruction, language-specific syllables such as [qy] should have been judged to be just as canonical as universally familiar syllables such as [ba]. For canonicity confidence judgements, coders were asked to attend to the same features that influenced their canonicity judgements.

However, the instructions to coders produced neither a sharp dissociation between the confidence measures (canonicity confidence or transcription confidence) and the utterance canonicity measure, nor between the two types of confidence measures. In fact the intercorrelations were statistically reliable in every case except between utterance canonicity and canonicity confidence, and even in that case the correlation showed a probability of less than .06 in both the 30-utterance and the 21-utterance analyses (see Tables III and V). Furthermore, both the confidence measures were highly correlated with transcription agreement.

The lack of sharp dissociation between the confidence measures and utterance canonicity suggests interrelations among the variables that we interpret as follows. The nearly statistically reliable correlation between canonicity confidence and utterance canonicity suggests that coders were more confident that they recognized canonicity when it was high than when it was low. The even higher correlation between transcription confidence and utterance canonicity suggests that coders may have been even more confident that their transcriptions were better

when utterances had been judged to be canonical than they were of the canonicity judgements themselves.

This pattern is consistent with our reasoning that one of the many issues influencing confidence judgements in transcription is canonicity itself. Transcribers have intuitive awareness of the well-formedness of speech—they recognize ill-pronounced utterances that might be the result of articulatory disorders, they notice the immaturity of infant vocalizations, and they identify non-speech sounds as being non-speech—all these abilities appear to be at least partly based on the recognition of canonicity. Thus, one of the various aspects of the confidence that coders may express regarding their transcriptions could be based on their awareness of canonicity.

The original motivation for the present research was an interest in the role of canonicity in transcription agreement, but coder confidence could play a particularly important part in establishing perspective on the results we obtain. One reason is that coder confidence may serve as a convenient proxy for a combination of many variables that influence transcription agreement. Of course, given that confidence levels have not been investigated in transcription agreement research in the past, the extent of the role of confidence levels as a proxy will surely require considerable research beyond that reported here. The present report provides a systematic first look at relations among canonicity, coder confidence and transcription agreement.

Relation of this research to inter-observer agreement in older children or mature speakers

While our empirical focus is on infant vocalizations, the information presented should, as suggested above, provide perspective on transcription agreement for other speaker populations as well. Not only are the immature characteristics of infant vocalizations found in the first months of life, but they are also heard (though less frequently) throughout early childhood (Oller, 2000a). Further, speakers with articulatory disorders or hearing impairment have been reported to produce sounds that violate principles of canonicity (Hudgins & Numbers, 1942; Weismer, 1984; Kent & Adams, 1989). Therefore, the significance of results on prediction of agreement in transcription and coding of infant vocalizations should encompass a much broader arena, providing both scientific and clinical insight on transcription reliability across many populations. In a similar way, it is possible that coder confidence may help broaden our perspective on intuitions of clinicians who are required to judge the quality of disordered speech.

That being said, there may be important differences between how clinicians and researchers approach transcription of different speaker populations. One could argue that generalization of these findings from one population to another is potentially risky given that the assumptions of the transcriber differ when transcribing pre-linguistic vocalizations in comparison to speech disorders, such as dysarthria or the speech of hearing-impaired individuals; the goals of transcription may differ. Our opinion is that, regardless of the assumptions of the transcriber, study of infant vocalizations is relevant to study of other speaker populations with respect to the ideas of perceptual templates and familiarity. Regardless of the assumptions of the transcriber, it is widely believed by perception theorists that there are always templates that guide perception. To the extent that a sample of speech to be transcribed has characteristics that are unfamiliar to the transcriber, whether through non-canonicity, non-nativeness, or vocal aberrations, there are reasons to expect perceptual variability and transcription disagreement. This of course does not imply that there would be no differences between outcomes across different populations—we merely contend that characteristics based on canonicity and other template-related factors should be operative for transcription agreement across samples from all populations of speakers.

Why not simply abandon phonetic transcription with infant sounds?

Given the fact that reliability is often low for transcription of infant samples, some have suggested simply abandoning its use with infants (Lynip, 1951). Our position is more moderate. We recognize the difficulties of using transcription, but we also recognize that there is an inevitable tendency to seek to characterize infant sounds in terms of the mature model of adult speech. This tendency applies to parents and scientists alike. Why does transcription continue to be utilized, and why do we think it *should* in some circumstances continue to be utilized, judiciously, in spite of reliability problems? The remaining paragraphs provide our answer.

The study of infant vocalization is crucial to understanding the development of human communication. Human language is overwhelmingly vocal, and throughout the first year of life the infant develops a rich repertoire of prelinguistic vocalizations that are used in primitively communicative ways (Oller, 1981; Stark, 1981; Stark, Bernstein, & Demorest, 1993). These vocalizations show a discernible relation with the phonetic characteristics of speech and are transformed through stages into mature well-formed syllables (Oller, 1980; Stark, 1980) that are then adapted into words (Vihman, 1996). Infant vocalizations prior to canonical babbling consist of acoustically very complex raw material, the apparent product of vocal exploration (Stark, 1981; Koopmans-van Beinum & van der Stelt, 1986; Papaeliou, Minadakis, & Cavouras, 2002). Such pre-canonical vocalizations, in spite of their complexity, can be categorized into identifiable groups as recognized by laboratory staff and parents: squeals, growls, raspberries, etc. In addition, there are well-recognized stages of vocal development, culminating in the controlled and repetitive production of canonical syllables in the second half year of life (Stark, 1981; Oller, 1995), each stage showing progressively greater approximation to the sounds of mature speech.

Parents and scientists face similar difficulty in interpretation of infant sounds. Early infant words appearing at the end of the first year, consisting of canonical syllables, clearly represent better approximations to adult words than pre-canonical sounds do, but they are still approximations. It is clear that these early word forms must be negotiated between parents and infants, because the infant sounds rarely if ever conform perfectly to adult models. It is not uncommon, for example, for a parent to ask “bottle?” or “ball?”, when the infant produces a syllable identified as [ba]. In such ways, it appears that parents and infants work out shared meanings through interaction based in part on a gross similarity of the utterances produced by the infant to potential words of the language to be learned. Since mature language is based on syllabic/segmental units, it can be argued that infant sounds *must* be interpreted by parents in this way, adapting them to the adult phonology to the extent possible, for the purposes of negotiation of shared meanings, a key process in early word learning. The more canonical vocalizations are, the more likely they are to be interpreted as words by parents (Papoušek, 1994).

The parent is thus *obliged* to use auditory judgement of the similarity between infant utterances and mature words in interaction with the infant, in spite of the fact that those judgements have limited inter-observer reliability (and limited construct validity). The auditory judgement represents a *phonetic* interpretation by the parent and it provides a reason that phonetic description of infant sounds based on auditory judgements is an inevitable aspect of the study of vocal development. Subjective phonetic interpretations of infant sounds by real listeners cannot be entirely supplanted in the study of vocal development by any presumed objective measure such as acoustic analysis. While it is definitely valuable to enhance our perspective on infant vocalization with acoustic measures and infra-phonologically-based coding, direct auditory phonetic interpretation of infant sounds will always come back as one kind of anchor for our research, because the infant’s learning, in the real interactions of word negotiation, hinges on it.

Thus, despite the questionability of using the IPA with infant vocalizations, there is continued interest in transcription as a tool for documentation of infant vocal production (see e.g. Stockman, Woods, & Tishman, 1981; MacNeilage & Davis, 1990; Vihman, 1996; Stoel-Gammon, 2001). The infant system must eventually take on a fully segmental form, and, as a consequence, any investigation of the emergence of language capability must ultimately (even if indirectly) be referenced to segment-like units. The roles that canonicity and confidence play in transcription of these units provides a perspective on how such auditory phonetic interpretations are made.

Acknowledgments

This work was supported by the National Institutes of Deafness and other Communication Disorders (R01DC006099, D. K. Oller PI and Eugene Buder Co-PI) and by the Plough Foundation.

References

- Aronson, E.; Ellsworth, PC.; Carlsmith, JM.; Gonzales, MH. Methods of research in social psychology. McGraw-Hill, Inc.; Boston, MA: 1990.
- Baranski JV, Petrusic WM. Probing the locus of confidence judgements: experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance* 1998;24:929–945. [PubMed: 9627426]
- Barrett S. The perceptual magnet effect is not specific to speech prototypes: new evidence from music categories. *Speech hearing and language: Work in progress* 1999;11:1–16.
- Bauer HR. The ethologic model of phonetic development: I. Phonetic contrast estimators. *Clinical Linguistics and Phonetics* 1988;2:347–380.
- Best CT, McRoberts GW. Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech* 2004;46:183–216. [PubMed: 14748444]
- Coussé, E.; Gillis, S.; Kloots, H.; Swerts, M. The influence of the labeller's regional background on phonetic transcriptions: implications for the evaluation of spoken language resources. In: Lino, M.; Xavier, M.; Ferreira, F.; Costa, R.; Silva, R., editors. *Proceedings of the 4th International Conference on Language Resources and Evaluation*; Paris: ELRA; 2004. p. 1447-1450.
- Cucchiari C. Assessing transcription agreement: methodological aspects. *Clinical Linguistics and Phonetics* 1996;10:131–155.
- Davis BL, MacNeilage PF. The articulatory basis of babbling. *Journal of Speech and Hearing Research* 1995;38:1199–1211. [PubMed: 8747814]
- Deutsch D. The tritone paradox: a link between music and speech. *Current Directions in Psychological Science* 1997;6:174–180.
- Duckworth M, Allen G, Hardcastle W, Ball M. Extensions to the International Phonetic Alphabet for the transcription of atypical speech. *Clinical Linguistics and Phonetics* 1990;4:273–280.
- Epstein W, Rock I. Perceptual set as an artifact of recency. *The American Journal of Psychology* 1960;73:214–228.
- Gibson JJ. A critical review of the concept of set in contemporary psychology. *Psychological Bulletin* 1941;38:781–817.
- Goldstone RL. Perceptual learning. *Annual Review of Psychology* 1998;49:585–612.
- Green, DM.; Swets, JA. Signal detection theory and psychophysics. Robert E. Krieger; Huntington, NY: 1966.
- Hudgins CV, Numbers FC. An investigation of intelligibility of speech of the deaf. *Genetic Psychology Monograph* 1942;25:289–392.
- Kearns KP, Simmons NN. Interobserver reliability and perceptual ratings: more than meets the ear. *Journal of Speech and Hearing Research* 1988;31:131–136. [PubMed: 3352250]
- Kent, RD.; Adams, SB. The concept and measurement of coordination in speech disorders. In: Wallace, SA., editor. *Perspectives on the coordination of movement*. Elsevier Science Publishers B.V.; North-Holland: 1989. p. 415-451.

- Kent, RD.; Read, C. Acoustic analysis of speech. Singular: Thomson Learning; Albany, NY: 2002.
- Kent RD, Kent JF, Weismer G, Martin RE, Sufit RL, Brooks BR, Rosenbek JC. Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics and Phonetics* 1989;3:347–358.
- Kirk, R. Introductory statistics. Brookes/Cole; Monterrey, CA: 1978.
- Koopmans-van Beinum, FJ.; van der Stelt, JM. Early stages in the development of speech movements. In: Lindblom, B.; Zetterstrom, R., editors. *Precursors of early speech*. Stockton Press; New York: 1986. p. 37-50.
- Kuhl, PK. Speech prototypes: studies on the nature, function, ontogeny and phylogeny of the “centers” of speech categories. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. *Speech perception, production and linguistic structure*. Ohmsha; Tokyo: 1992. p. 240-264.
- Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 1992;255:606–608. [PubMed: 1736364]
- Lieberman AM, Delattre PC, Gerstman LJ, Cooper FS. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology* 1956;52:127–137. [PubMed: 13345983]
- Louko LJ, Edwards ML. Issues in collecting and transcribing speech samples. *Topics in Language Disorders* 2001;21:1–11.
- Lu Z-L, Doshier BA. Perceptual learning retunes the perceptual template in foveal orientation identification. *Journal of Vision* 2004;4:44–56. [PubMed: 14995898]
- Lynch MP, Oller DK, Steffens ML, Levine SL. Onset of speech-like vocalizations in infants with Down syndrome. *American Journal on Mental Retardation* 1995;100:68–86. [PubMed: 7546639]
- Lynip A. The use of magnetic devices in the collection and analyses of the preverbal utterances of an infant. *Genetic Psychology Monographs* 1951;44:221–262. [PubMed: 14887926]
- MacNeilage, PF.; Davis, BL. Acquisition of speech production: the achievement of segmental independence. In: Hardcastel, WJ.; Marchal, A., editors. *Speech production and speech modelling*. Kluwer; Dordrecht: 1990. p. 55-68.
- MacNeilage PF, Davis BL. On the origin of internal structure of word forms. *Science* 2000;288:527–531. [PubMed: 10775113]
- Marbach JJ, Franks A, Raphael KG, Janal MN, Hirschhorn-Roth R. Reliability of clinician judgements of bruxism. *Journal of Oral Rehabilitation* 2003;30:113–118. [PubMed: 12535135]
- McCarthy JJ. Feature geometry and dependency: a review. *Phonetica* 1988;45:84–108.
- McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature* 1976;264:746–748. [PubMed: 1012311]
- Nathani S, Ertmer DJ, Stark RE. Assessing vocal development in infants and toddlers. *Clinical Linguistics and Phonetics* 2006;20:351–369. [PubMed: 16728333]
- Oller DK. Infant vocalization and the development of speech. *Allied Health and Behavioral Sciences* 1978;1:523–549.
- Oller, DK. The emergence of the sounds of speech in infancy. In: Yeni-Komshian, G.; Kavanagh, J.; Ferguson, C., editors. *Child phonology, Volume 1, Production*. Academic Press; New York: 1980. p. 93-112.
- Oller, DK. Infant vocalizations: exploration and reflexivity. In: Stark, RE., editor. *Language behavior in infancy and early childhood*. Elsevier North Holland; New York: 1981. p. 85-104.
- Oller, DK. Development of vocalizations in infancy. In: Winitz, H., editor. *Human communication and its disorders: A review*. Vol. IV. York Press, Inc.; Timonium, MD: 1995. p. 1-30.
- Oller, DK. The emergence of the speech capacity. Lawrence Erlbaum and Associates; Mahwah, NJ: 2000a.
- Oller, DK. Transcription and categorization of infant vocalization. *International Child Phonology Conference*; Cedar Rapids, IA: 2000b.
- Oller, DK.; Delgado, RE. *Logical International Phonetics Programs (Version Windows)*. Intelligent Hearing Systems Corp.; Miami: 1999.
- Oller DK, Eilers RE. Phonetic expectation and transcription validity. *Phonetica* 1975;31:288–304. [PubMed: 1124265]

- Oller DK, Ramsdell HL. A weighted reliability measure for phonetic transcription. *Journal of Speech and Hearing Research* 2006;49:1391–1411.
- Oller DK, Eilers RE, Basinger D. Intuitive identification of infant vocal sounds by parents. *Developmental Science* 2001;4:49–60.
- Oller DK, Eilers RE, Bull DH, Carney AE. Pre-speech vocalizations of a deaf infant: a comparison with normal metaphonological development. *Journal of Speech and Hearing Research* 1985;28:47–63. [PubMed: 3981997]
- Oller DK, Eilers RE, Steffens ML, Lynch MP, Urbano R. Speech-like vocalizations in infancy: an evaluation of potential risk factors. *Journal of Child Language* 1994;21:33–58. [PubMed: 8006094]
- Oller DK, Wieman L, Doyle W, Ross C. Infant babbling and speech. *Journal of Child Language* 1975;3:1–11.
- Papaeliou C, Minadakis G, Cavouras D. Acoustic patterns of infant vocalizations expressing emotions and communicative functions. *Journal of Speech, Language, and Hearing Research* 2002;45:311–317.
- Papoušek, M. Vom ersten Schrei zum ersten Wort: Anfänge der Sprachentwicklung in der vorsprachlichen Kommunikation. Verlag Hans Huber; Bern: 1994.
- Pollack, I.; Decker, LR. Confidence ratings, message reception, and the receiver operating characteristic. In: Swets, JA., editor. *Signal detection and recognition by human observers*. John Wiley and Sons; New York: 1964. p. 592–608.
- Pollock KE, Berni MC. Transcription of vowels. *Topics in Language Disorders* 2001;21:22–40.
- Ralston JV, Johnson K. Context effects as a function of perceptual set. *Journal of the Acoustical Society of America* 1990;87:S46–S47.
- Richards JS, Hicken BL, Putzke JD, Ness T, Kezar L. Reliability characteristics of the Donovan spinal cord injury pain classification system. *Archives of Physical Medicine and Rehabilitation* 2002;83:1290–1294. [PubMed: 12235610]
- Robb MP, Carcace AT. Estimation of formant frequencies in infant cry. *International Journal of Pediatric Otorhinolaryngology* 1995;32:57–67. [PubMed: 7607821]
- Roberts WT, Higham PA. Selecting accurate statements from the cognitive interview using confidence ratings. *Journal of Experimental Psychology: Applied* 2002;8:33–43. [PubMed: 12009175]
- Robinson MD, Johnson JT. Recall memory, recognition memory, and the eyewitness confidence-accuracy correlation. *Journal of Applied Psychology* 1996;81:587–594.
- Shriberg LD, Lof GL. Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics* 1991;5:225–279.
- Stark RE. Features of infant sounds: the emergence of cooing. *Journal of Child Language* 1978;5:379–390. [PubMed: 701415]
- Stark, RE. Stages of speech development in the first year of life. In: Komshian, GY.; Kavanagh, J.; Ferguson, C., editors. *Child phonology, Volume 1, Production*. Academic Press; New York: 1980. p. 73–90.
- Stark R. Infant vocalization: a comprehensive view. *Infant Medical Health Journal* 1981;2:118–128.
- Stark R, Bernstein LE, Demorest ME. Vocal communication in the first 18 months of life. *Journal of Speech and Hearing Research* 1993;36:548–558. [PubMed: 8331912]
- Stockman IJ, Woods DR, Tishman A. Listener agreement on phonetic segments in early infant vocalizations. *Journal of Psycholinguistic Research* 1981;10:593–617. [PubMed: 7343665]
- Stoel-Gammon C. Transcribing the speech of young children. *Topics in Language Disorders* 2001;21:12–21.
- Swets, JA. *Signal detection and recognition by human observers*. John Wiley and Sons; New York: 1964.
- Van Zandt T. ROC curves and confidence judgements in recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition* 2000;26:582–600.
- Vihman, MM. *Phonological development: The origins of language in the child*. Blackwell Publishers; Cambridge, MA: 1996.
- Warren RM, Obusek CJ. Speech perception and phonemic restoration. *Perception and Psychophysics* 1971;9:358–362.

- Weismer, G. Acoustic descriptions of dysarthric speech: perceptual correlates and physiological inferences. In: Rosenbek, JC., editor. Current views of dysarthria: Nature, assessment and treatment. Seminars in Speech and Language; New York: Thieme-Stratton; 1984. p. 293-314.
- Weismer, G.; Martin, RE. Acoustic and perceptual approaches to the study of intelligibility. In: Ball, MJ.; Kent, RD., editors. Intelligibility in speech disorders. John Benjamins Publishing Company; Philadelphia, PA: 1992. p. 67-118.
- Wells, GA.; Murray, DM. Eyewitness confidence. In: Wells, GA.; Loftus, EA., editors. Eyewitness testimony: psychological perspectives. Cambridge University Press; Cambridge, UK: 1984.
- Werker JF, Tees RC. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 1984;7:49–63.

Figure 1. Waveform and spectrographic display of two syllables from the primary sample. Spectrogram (a) shows a canonical syllable, with cursors surrounding the timely formant transitions from C to V segments of approximately 68 ms. Spectrogram (b) shows a non-canonical syllable with protracted formant transitions of ~147 ms.

Figure 2.

(a) Transcription agreement for non-canonical and canonical utterances from the primary sample. (b) Transcription agreement for non-native sounds and native sounds. (c) Transcription agreement for utterances with aberrant vocal quality vs those with typical vocal quality.

Figure 3.

Overall weighted transcription agreement for non-canonical and canonical utterances for each coder comparison on transcription of the supplementary sample.

Table I

Six levels of coding designated in the coding protocol. Transcribers coded various aspects of the utterances in levels 1, 3, and 5, while making confidence judgements on the accuracy of each code in levels 2, 4, and 6.

Levels of coding				
(1) Structure judgement	N	C	N	N
(2) Structure confidence	5	4	3	5
(3) Canonicity judgement		*	*	
(4) Canonicity confidence	4	5	3	5
(5) Transcription	a	υ	i	ə
(6) Transcription confidence	4	3	5	6

* indicates that a structure is non-canonical.

Table II

Example of the alignment of one utterance across the 8 coders in the primary sample. Column 1 lists the coders, column 2 presents the transcription alignments, and column 3 the structure and canonicity alignments. The phonetic symbols aligned in column 2 were maximally matched based on featural similarity to ensure minimal discrepancy across codes.

Compared transcription			
Coder	Transcriptions		Structure and canonicity
Standard coder	ε	υ	m N* N C
Comparator 1	Λ		m N C
Comparator 2	ə		m N* C*
Comparator 3	Λ		m N* C*
Comparator 4	Λ		m N* C*
Comparator 5	ə		m N C
Comparator 6		α	m N* C*
Comparator 7	Λ		m N* C*

* indicates that a structure is non-canonical.

Table III

Means, standard deviations, and correlations for the 30-utterance sample, with utterance canonicity, canonicity confidence and transcription confidence regressed on the weighted transcription agreement values.

	Weighted transcription agreement (sig.)	Utterance canonicity (sig.)	Canonicity confidence (sig.)	Transcription confidence
Weighted transcription agreement	1			
Utterance canonicity	.376 (p=.041)	1		
Canonicity confidence	.527 (p=.003)	.351 (p=.057)	1	
Transcription confidence	.566 (p<.001)	.671 (p<.000)	.560 (p<.001)	1
M	.598	.532	5.387	4.552
SD	.151	.284	.559	.748

Table IV

Regression results for utterance canonicity, canonicity confidence, and transcription confidence regressed on the weighted transcription agreement values for the 30-utterance sample.

Independent Variables	b	β	t	Sig.
<i>Model 1</i>				
Utterance canonicity	.200	.376	2.148	.041
<i>Model 2</i>				
Utterance canonicity	.004	.007	.032	.975
Canonicity confidence	.083	.306	1.645	.112
Transcription confidence	.079	.391	1.665	.108

Model 1 R-Square=.141; Model 2 R-Square=.385; R-Square change=.243, Sig. F change=.013.

Table V

Means, standard deviations, and correlations for the 21-utterance sample, with utterance canonicity, canonicity confidence, and transcription confidence regressed on the weighted transcription agreement values.

	Weighted transcription agreement (sig.)	Utterance canonicity (sig.)	Canonicity confidence (sig.)	Transcription confidence
Weighted transcription agreement	1			
Utterance canonicity	.626 (p=.002)	1		
Canonicity confidence	.700 (p<.000)	.430 (p<.052)	1	
Transcription confidence	.645 (p=.002)	.528 (p=.014)	.768 (p<.000)	1
M	.562	.592	5.361	4.547
SD	.150	.272	.512	.715

Table VI

Regression results for utterance canonicity, canonicity confidence, and transcription confidence regressed on the weighted transcription agreement values for the 21-utterance sample.

Independent Variables	b	β	t	Sig.
<i>Model 1</i>				
Utterance canonicity	.346	-.626	3.499	.002
<i>Model 2</i>				
Utterance canonicity	.210	.380	2.160	.045
Canonicity confidence	.132	.478	2.050	.056
Transcription confidence	.016	.078	.313	.758

Model 1 R-Square=.392; Model 2 R-Square=.622; R-Square change=.230, Sig. F change=.18.

Table VII

Example of common vowel and consonant disagreements found across transcriptions.

Disagreement type	Examples*	Percentage
consonant – consonant	b ↔ p	16.9%
	ɣ ↔ h	8.0%
	d ↔ z/v ↔ w	4.5% of each error type
vowel – vowel	Λ ↔ ə	15.5%
	a ↔ ɑ	11.8%
	a ↔ æ	7.6%
Vowel – consonant	j ↔ I	10.8%
	ʊ ↔ ə	8.1%
	ʊ ↔ v/ɥ ↔ a	5.4% of each error type

* ↔ indicates that the two phonetic symbols were used to represent the same sound heard.