The Origins of Phonology and Lexicon in Infancy (OPAL): Phonological Abstraction Before Perceptual Attunement?

Denis Burnham[1], Catherine Best[1], Antonia Götz[1], Marina Kalashnikova[2,1], Elizabeth Johnson[3], Eylem Altuntas[1], Anne Cutler[1,4,*]

[1]MARCS Institute for Brain, Behaviour & Development, Western Sydney University;
[2]Basque Center on Cognition, Brain & Language [3]Child Language & Speech Studies Lab, University of Toronto; [4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
denis.burnham@westernsydney.edu.au; c.best@westernsydney.edu.au; a.goetz@westernsydney.edu.au;
m.kalashnikova@bcbl.eu; elizabeth.johnson@utoronto.ca;

Inspired by, and dedicated to the memory of, Distinguished Professor Anne Cutler

ABSTRACT

Recently it was found that Dutch-speaking adults adopted from Korea at less than 6 months learned to perceive a Korean consonant contrast that is non-native in Dutch better than native Dutch speakers lacking Korean experience. They also better generalised that perceptual learning to other articulation places and to speech production. The inference that phonological abstraction occurs before 6 months, prior to language-specific attunement (6-10 months), begs direct evidence from infants. Our project, ‘Origins of Phonology and Lexicon’ (OPAL), is designed to address that lacuna. In this paper we (i) derive and describe tasks for evaluating three types of phonological abstraction: Protoword Structure, Phonological Inventories, and Phoneme Features; (ii) describe our method for testing these in infancy; and (iii) validate these tasks with adults. While adults showed differences among tasks, they were able to perform each abstraction with just one feedback trial, providing a firm basis for our ongoing infant studies.

Keywords: phonological abstraction, infant speech perception, perceptual attunement, adult validation

1. INTRODUCTION

1.1. Background

In recent studies by Choi, Cutler and colleagues, a group of Dutch-speaking adults who had been adopted from Korea as infants between 3 and 6 months, with no further experience of Korean, were trained to identify the Korean fortis/lenis/aspirated alveolar stops /t*/-/t/-/t^h/, which are not phonemic in Dutch [1, 2]. The adoptees learned the distinction more quickly than a control group of native Dutch-speaking adults lacking any Korean language experience. While this might suggest that the adoptees had stored a complete compilation of the Korean phoneme repertoire, it was also found that the adoptees surpassed the controls in (i) generalising from the learned alveolar contrast to the same fortis/lenis/aspirated contrast in bilabial and velar places, (ii) producing these Korean consonants, and (iii) showing higher speech perception-production correlations. Together these results suggest that what had been gleaned before the age of 6 months was not mere memory for the Korean consonant inventory, but rather phonologically abstract representations, in this case the contrasting fortis/lenis/aspirated features of stop consonants. Consequently, phonological knowledge appears to be gained before the emergence of perceptual attunement, the process by which infants tune in to the phonemes and contrasts of their ambient language between about 6 to 10 months [4].

These adult results beg the questions of how and when abstract phonological representations are laid down in infancy and this is what our ‘Origins of Phonology and Lexicon’ (OPAL) project has been funded to investigate [3]. This paper concerns the development (i) of tasks for investigating the timing, nature and extent of phonological abstraction in infancy, (ii) of methods for testing performance on the tasks; and (iii) validating the tasks and measures with adult participants.

1.2. Development of Phonological Abstraction Tasks

A recent EEG study has provided a proof of concept for a method to investigate phonological abstraction by 5-month-old infants [5]. Infants were familiarised with two classes of spoken non-words differing only in syllable location, viz, ABA (e.g., *baluba. gotigo, etc.*), paired with an arbitrary visual cartoon, e.g., a fish, and AAB (e.g., *babalu, gogoti, etc.*) paired with another visual cartoon, e.g., a lion. In the subsequent test trials with new non-words of each syllable structure, infants showed neural mismatches to incongruent word-label pairings (i.e., reverse pairings to those in familiarisation), but not to congruent pairings (same as familiarisation), suggesting that infants had abstracted the opposing multi-syllabic structures, ABA vs AAB. We have adapted this method to test three types of phonological abstraction in infants under 6 months. Derivation of these three phonological abstraction tasks, and determination of familiarisation and test stimuli, are set out below.

1.2.1. Protoword Structure (PRO)

Newborns discriminate differences in the structure of syllable sequences, e.g., AAB (*babamu, gegeba*) vs ABC (*bamuge, gebamu*) [6,7]. In addition to discriminating such structural differences, young infants can also detect changes in position of a syllable within a series (A*B*A→AA*B*) as in [5] and in statistical learning studies with artificial grammars [8]. Our protoword structure task (Table 1) goes beyond syllable sequences in *words* to test sensitivity to phonotactics, i.e., phoneme position in *syllables*.

Phase & Phonemes	**Category A CCV**	**Category B CVC**
Learning 3-syll sets from:	/blV, glV, spV, stV, ʃmV/	/bVl, gVl, sVp, sVt, ʃVm/
example:	*blee-glar-spoo*	*beel-garl-soop*
Test	As for Learning but new nonwords	
example:	*blar-gloo-spee*	*barl-gool-seep*

Table 1: Proto-Word Structures (PRO) Task (V = vowel).

1.2.2. Phonological Inventories (INV)

Infant sensitivity to language differences in prosodic rhythms is supported by French newborns' successful discrimination of stress-timed English from mora-timed Japanese but not from stress-timed Dutch [9]. It has been proposed that rhythm differences between languages reflect differences in proportions of vowels and consonants [10]. Moreover, differences in consonant/vowel distributions have consequences for the phonological abstraction of permissible word structures [11]. Our phonological inventories task (Table 2) will provide the first test of infants' ability to distinguish among consonant/vowel distributions.

Phase & Phonemes	**Category A 2V-8C**	**Category B 8V-2C**
Learning 3-CV words from:	C: /b, k, dʒ, s, l, w, m, n/ V: /iː, æɔ/	C: /m, k/ V: /iː, ɐː, oː, ʉː, oɪ, æɪ, æɔ, ɑe/
example:	*bee-kou-jee-sou*	*mee-koy-mar-kay*
Test	As for Learning but new nonwords	
example:	*bou-kee-jou-see*	*moy-kar-may-zor*

Table 2: Phonological Inventories (INV) Task (V = vowel, C = consonant)

1.2.3. Phoneme Features (FEA)

Between 14 and 19 months infants develop phonological categories that: are durable across changes in English accents, mature over age, and are correlated with infants' vocabulary size [12,13,14].

We will investigate a possible precursor of phonological category formation in infancy – assessing whether infants can categorise consonants by a phonological feature contrast akin to that tested by Choi, Cutler and colleagues [1, 2], in this case voiced vs voiceless consonants (see Table 3).

Phase & Phonemes	**Category A Voiced**	**Category B Voiceless**
Learning 3-syll words from:	/bV, vV, dʒV/	/pV, fV, tʃV/
example:	*bee-var-jor*	*par-for-choo*
Test*	/dV, zV, gV/	/tʃV, sV, kV/
example	*dor-zoo-gae*	*too-sae-kee*

Table 3: Phoneme Features (FEA) Task (V = vowel). *Test phase uses new consonants, to test feature generalisation.

1.2.4. Summary

As can be seen the three tasks are highly complex, requiring us to calculate the size of the expected effects in order to determine sample sizes and analysis models for the infant experiments. Therefore, we first needed to validate the three tasks with adults, which is what we report in this paper.

2. METHOD[i]

2.1 Design

A within-subject design was employed: 3 tasks (Protoword Structure / Phonological Inventories / Phonemic Features) x 2 phases (Learning / Test).

2.2 Participants

65 adults participated but 11 had incomplete data. 54 data sets were included for trials to criterion (TTC), and 53 for reaction times (RT) (one had missing RT data in one cell). Language background varied within the constraint that each participant must be proficient in English (18 L1-English [8 monolingual]; 36 L2-English). Mean age was 38.54 years (median = 34.5, range = 17-77, standard deviation = 14.21).

2.3. Experimental Environment

The software for the three tasks was developed in PsychoPy, a Python-based package for behavioural studies [15]. The study was conducted online (due to the COVID pandemic) using Pavlovia, an extension of PsychoPy [16]. All instructions were in English.

2.4. Stimulus Materials

For each task, two contrasting categories (A, B) of multi-syllabic non-words were created, composed of CCV vs CVC syllables (C=consonant, V=vowel) for *Protoword Structure* (PRO); 2V-8C vs 8V-2C inventories for *Phonological Inventories* (INV); and voiced (Vd) vs voiceless (Vl) initial consonants for *Phonemic Features* (FEA) (see Tables 1-3). For each task, 480 words were created, 240 each for Category A and B. Words were recorded by a different female Australian English speaker for each task.

2.5. Procedure

In each task within a two-alternative forced-choice implicit learning task, participants heard non-words from categories (A or B) and were required to identify the category of each word by pressing the A computer key for Category 'A', and the 'L' key for Category B. There were two phases, Learning and Test. Each drew stimuli from 120 Category A and 120 Category B words selected randomly without replacement from the relevant category. No more than three words from either category appeared in succession.

At the start of the Learning phase participants heard, 'This sound belongs to A', followed by a word from List A. This was the *only* trial with feedback; *all* remaining trials in Learning and Test lacked feedback. The sequence in ensuing trials was: (i) 'Ready' on the screen for 0.5secs; (ii) auditory word presented; (iii) a required key press ('A' or 'L'); (iv) next trial after response or after a time-out period of 1.5secs. No-response trials were not repeated. Key press responses and reaction times were recorded.

Participants proceeded from Learning to Test phase, and from Test phase to task-end if they (i) made 7 correct responses in a moving window of 8 trials (binomial probability = .03516, p<.05), *or* (ii) if they did not reach this criterion within 50 trials. Between Learning and Test phases, and between the Test phase of one task and the Learning phase of the next, there was a break and participants advanced via pressing the space bar. All participants completed all three tasks, and task order was counterbalanced.

3. RESULTS

The dependent measures were Response Criterion, and Reaction Times for correct responses in both Learning and Test phase in each task.

3.1. Response Criterion Measurements

3.1.1 Participants Reaching Criterion

The number of participants reaching criterion (7 correct responses in a moving window of 8 trials within the maximum of 50 trials) in each phase of each task are shown in the upper row of Table 4. A 2 x 3 chi-square analysis found no differences in number of participants reaching criterion across tasks or phases, χ^2 (2) = 0.22, p > .05. Further chi-square tests for each phase in each task were also not significant.

See the lower row of Table 4 for the binomial probability of reaching criterion against chance (0.5). In both phases of the PRO and INV tasks, the number of participants (/54) reaching criterion was greater than chance. But in neither FEA phase did the number reaching criterion exceed chance.[ii]

3.1.1 Trials to Criterion (TTC)

Figure 1 (checkered plots) and Table 4 upper row) show, for each task, mean trials to criterion for participants reaching criterion within the maximum of 50 trials per phase. Those who did not reach criterion were given a score of 57 (maximum number of trials (50) + the minimum number to reach criterion thereafter (7) see solid bars in Figure 1). This latter 'All Participants' measure affords repeated measures analysis as each of the 54 participants has a score for each of the six task x phase cells.

	PRO		INV		FEA	
	Learn	Test	Learn	Test	Learn	Test
n	35	33	34	34	26	29
p =	.014*	.049*	.027*	.027*	.608	.292

Table 4: n = number (of N= 54) per task and phase reaching criterion within 50 trials. p = binomial probability the proportion is above chance (.5). *p <.05.

A 3-task x 2-phase analysis of variance (ANOVA) of All Participants revealed no significant main effect for phase, F(1,53) = 0.05, p > .05, or its interactions. As FEA was the only task that incorporated (i) categories distinguished by a language-specific phonological feature distinction (voicing) and (ii) a

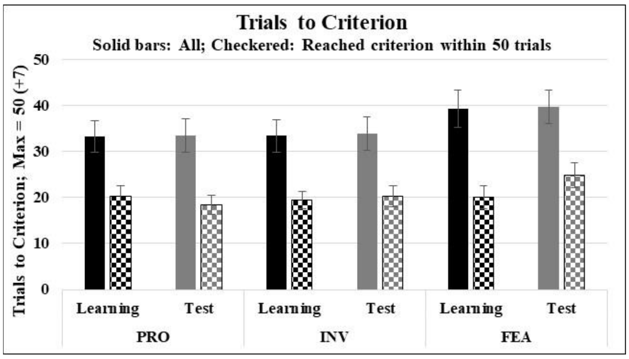

Figure 1: Trials to Criterion (TTC) for all participants in 50 (+7) trials and for only those reaching criterion within 50 trials. Error bars are standard errors of the mean. (s.e.m.)

generalisation from Learning to Test Phases, we ran planned comparisons to compare (i) FEA to the other two tasks combined (FEA vs PRO+INV), and (ii) the other two tasks (PRO vs INV), both of which had neither a feature distinction nor generalisation from Learning to Test. There were significantly more TTCs for FEA than for the mean of PRO and INV, F(1,53) = 5.39, p < .05, but no significant difference between PRO and INV, F(1,53) = 0.01, p > .05.[iii]

3.3.2. Mean Reaction Times (RTs)

Reaction Times (RTs) cannot be less than zero, so they are typically skewed. To correct for this, $\log^{10}$ values were calculated. These were used in the analyses and are plotted in Figure 2.

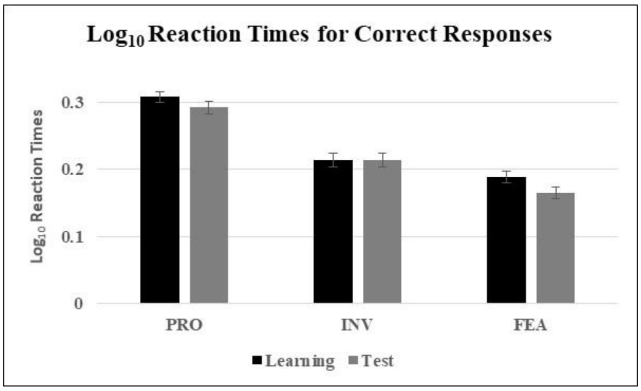

Figure 2: Mean $\log^{10}$ Reaction Times for Learning and Test phases in each of the three tasks. Error bars are standard errors of the mean.

A 3-task x 2-phase ANOVA revealed that RTs decreased significantly from Learning to Test across tasks, F(1, 52) = 11.34, p < .01. There was a main effect of task, indicating that RTs were significantly shorter for FEA than for PRO and INV combined, F(1, 52) = 98.81, p < .001. In turn, RTs were significantly shorter for INV than PRO, F(1, 52) = 255.89, p < .001.[iv]

4. DISCUSSION

Three phonological abstraction tasks – *Protoword Structure* (PRO), *Phonological Inventories* (INV), *Phonetic Features* (FEA) – were developed and validated with adults with a view to using these in tests of phonological abstraction in infancy.

Adults engaged in all three tasks, but the FEA task appears to be more difficult than either the PRO or the INV tasks because: (i) the number of participants reaching criterion within 50 trials was not significantly above chance in either Learning or Test for FEA (in PRO and INV it was significant in both phases) and (ii) there were more TTCs in FEA than in PRO and INV. Inspection of TTC data for only those reaching criterion (Figure 1) suggests an increase in TTC for FEA from Learning to the (Generalisation) Test. However, this observation is weakened by the fact that (i) this comparison was not significant in the All Participants TTC ANOVA, (ii) in neither FEA Learning nor Test was the number of participants reaching criterion greater than chance, and (iii) the faster RTs in FEA than for PRO or INV occurred irrespective of phase.

In contrast to the TTC analyses, RTs were surprisingly significantly shorter for FEA than PRO and INV, which may seem to imply, somewhat paradoxically, that FEA was an easier task. However, the TTC analyses imply that FEA was more difficult (were significantly greater TTCs than PRO ad INV), so it appears that participants may have responded more quickly in the FEA task simply because they had given up trying to figure out the basis of the two categories.

It is of passing interest that there was no effect of language background (L1-English vs L2-English) on either TTC or RT (see endnotes ii, iii, iv). There is a possible effect of language background on numbers reaching criterion in the FEA task (due to the use of an English voicing contrast), but the only firm conclusion that can be drawn is that, in general, the results here are robust over L1-English and L2-English adults.

Despite differences in FEA versus PRO and INV, the adults were able to complete all three tasks under difficult conditions, i.e., there was feedback on only one preliminary trial and that feedback was for only one of the two categories. In addition, they were given no information regarding the basis of the categories, which is, itself, the same constraint that the infants will face. So, together the results augur well for using these tasks (in EEG adaptations) with infants.

5. SUMMARY AND CONCLUSIONS

Three tests of phonological abstraction were developed and were validated with adults. One task, FEA, was more difficult that the other two, PRO and INV, but adults were able to engage with all tasks under very difficult conditions that would also be the case for infants. The results provide a solid foundation for the infant studies in this project.

6. APPENDIX – FURTHER DATA

In addition to the three tasks here, the OPAL project also includes a fourth task – a double test of phonological abstraction. In this the two categories to be learned are based on a consonantal place of articulation contrast, *plus* an additional learning-to-test change of modality (auditory-only to visual-only or visual-only to auditory-only). Preliminary infant data for this task is reported in ICPhS 2023 paper 42.

7. REFERENCES

[1] Choi, J., Broersma, M. and Cutler, A., 2017. Early phonology revealed by international adoptees' birth language retention. *Proceedings of the National Academy of Sciences 114*, pp.7307-7312.

[2] Choi, J., Cutler, A. and Broersma, M., 2017. Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory. *Royal Society Open Science 4*, 60660.

[3] Burnham, D., Best, C. (Cutler, A., Johnson, E., Kalashnikova, M.) Origins of Phonology and Lexicon: Abstract representations before 6 months. Australian Research Council Discovery Grant, DP00025567.

[4] Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology 50*, pp.509-535.

[5] Kabdebon, C. and Dehaene-Lambertz, G., 2019. Symbolic labeling in 5-month-old human infants. *Proceedings of the National Academy of Sciences*, *116*(12), pp. 5805-5810.

[6] Gervain, J., Macagno, F., Cogoi, S., Peña, M. and Mehler, J., 2008. The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, *105*(37), pp.14222-14227.

[7] Gervain, J., Berent, I. and Werker, J.F., 2012. Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, *24*(3), pp.564-574.

[8] Saffran, J.R., Aslin, R.N. and Newport, E.L., 1996. Statistical learning by 8-month-old infants. *Science*, *274*(5294), pp.1926-1928.

[9] Nazzi, T., Bertoncini, J., & Mehler, J. 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), pp. 756

[10] Ramus, F., Nespor, M. and Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*,*73*, pp.265-292.

[11] Nazzi, T., & Cutler, A. 2019. How consonants and vowels shape spoken-language recognition. *Annual Review of Linguistics*, *5*, pp. 25-47.

[12] Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. 2009. Development of phonological constancy: Toddlers' perception of native-and Jamaican-accented words. *Psychological Science*, *20*(5), pp. 539-542.

[13] Mulak, K.E., Best, C.T., Tyler, M.D., Kitamura, C. and Bundgaard-Nielsen, R.L., 2010. Vocabulary size predicts the development of phonological constancy: An eyetracking study of word identification in a non-native dialect by 15-and 19-month-olds. In *Proceedings of the 20th International Congress on Acoustics*.

[14] Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., & Irwin, J. R. 2013. Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development*, *84*(6), pp. 2064-2078.

[15] Peirce, J.W., 2007. PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, *162*(1-2), pp.8-13.

[16] Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J.K., 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), pp.195-203.

[i] We acknowledge assistance with *Programming:* Dr Leidy Castro-Meneses, Johnson Chen, Dr Benjawan Kasisopa, Prof Ronan Reilly, Dr Chris Wang, Mary Youkhanna; *Design*: Dr. Pelle Söderström; *Stimulus construction*: Dr Benjawan Kasisopa, Dr Chris Wang; *Stimulus voices*: Dr Ruth Brookman, Caitlin Hooper, Kate Jones; *Collation, stimuli & data*: Michael Barrett, Nur Hazwani Abd Hadi, Supreet Kaur; *Recruiting*: Jessica Simpson.

[ii] See the following table for the percentage of L1-English and L2-English participants reaching criterion. Low and unequal numbers preclude firm conclusions, but we note the following differences:

- PRO: similar percentages
- INV: more L2-English than L1-English participants reached criterion (maybe due to the more diverse language experience of L2 speakers)
- FEA: more L1-English than L2-English reached criterion (maybe due to the voicing contrast being based on an English phonetic voicing distinction).

English status	**N**	**PRO**		**INV**		**FEA**	
		Lrn	**Tst**	**Lrn**	**Tst**	**Lrn**	**Tst**
L1-Eng	**18**	61	61	50	39	50	67
L2-Eng	**36**	64	61	69	75	44	47

[iii] Further analyses of TTC for L1-English vs L2-English revealed no significant interactions of either Phase or Task with English status (L1, L2), so the results of the main analysis – more TTC for FEA than for the mean of PRO and INV – hold across both language background groups.

[iv] Further analyses of log RTs for L1-English vs L2-English revealed no significant interactions of Phase or Task with English status (L1, L2), so the results of the main analysis – increased RTs from Learning to Test, FEA with shorter RTs than PRO and INV, and INV with shorter RTs than PRO – hold across both language background groups.