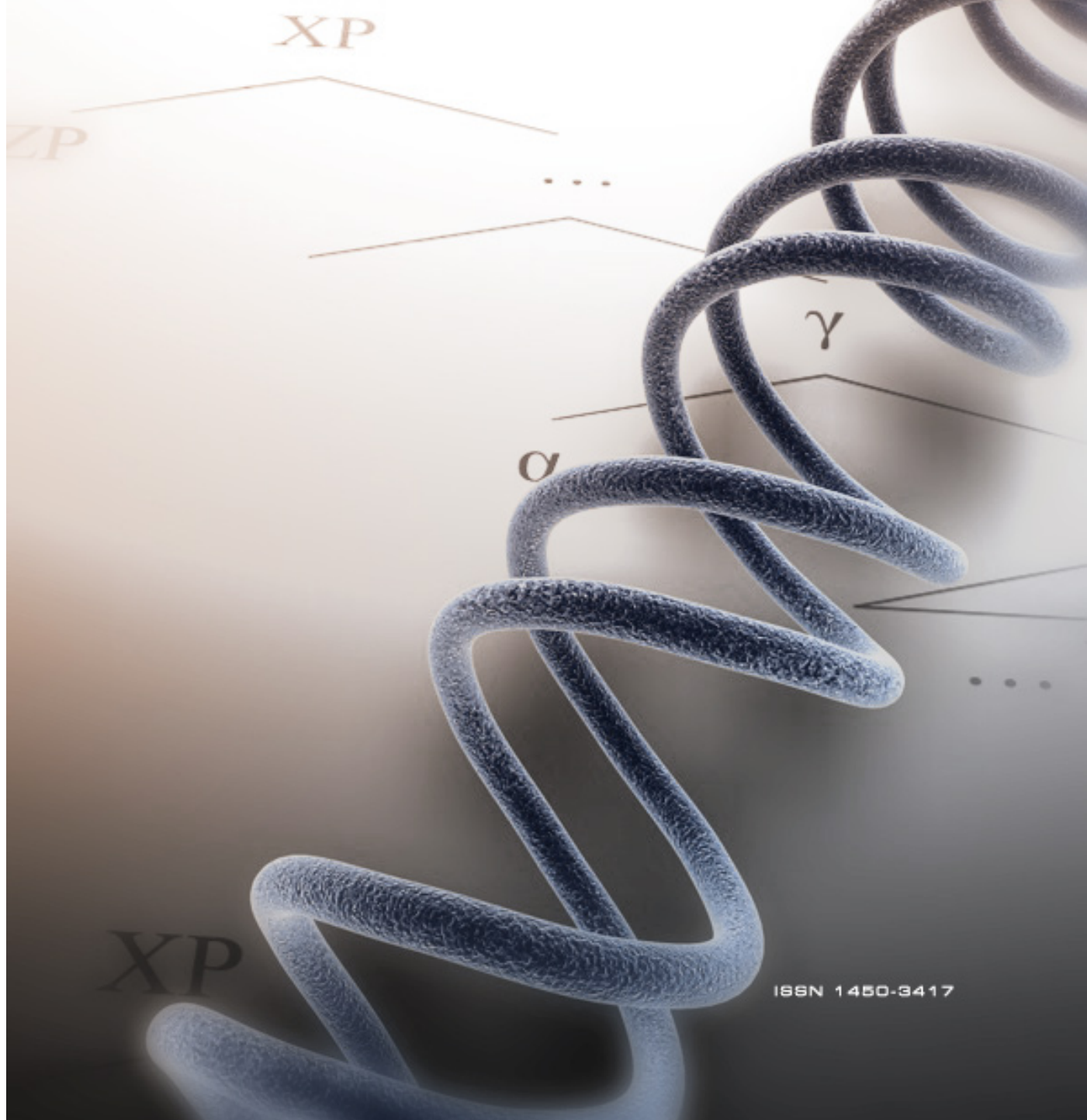




# biolinguistics

volume 14

2020



ISSN 1450-3417

# Attention to People Like You: A Proposal Regarding Neuroendocrine Effects on Linguistic Variation

Míša Hejná<sup>1,\*</sup>, Lauren Ackerman<sup>2</sup>, & Joel C. Wallenberg<sup>3</sup>

<sup>1</sup> Department of English, Faculty of Arts, Aarhus University, Denmark

<sup>2</sup> School of Education, Communication, & Language Science, Faculty of Humanities & Social Sciences, Newcastle University, United Kingdom

<sup>3</sup> School of English Literature, Language, & Linguistics, Faculty of Humanities & Social Sciences, Newcastle University, United Kingdom

\* Corresponding author: [misa.hejna@cc.au.dk](mailto:misa.hejna@cc.au.dk)



MH: <https://orcid.org/0000-0002-9328-2603>

LA: <https://orcid.org/0000-0002-3223-8641>

JCW: <https://orcid.org/0000-0002-1837-6765>

Although the literature on language change has often replicated and discussed a pattern in which female speakers lead in changes that occur below the level of awareness, there is no consensus on why this pattern should arise. Interestingly, recent findings in endocrinology show that differences in prenatal testosterone exposure can impact learning patterns. In the light of these findings, we first present preliminary results consistent with the hypothesis that a biological factor, prenatal exposure to androgens, can have a small, continuous biasing effect on linguistic variation, namely the variable duration of pre-aspiration conditioned by voiceless obstruents in Tyneside English. Second, we propose an explanatory model in which the biological factor—prenatal testosterone exposure—creates subtle bias in how speakers learn linguistic variants and suggest that some reported sex effects are derivative. This model is compatible with the high tendency for females to lead in language change from below (Labov 1990: 206).

*Keywords:* pre-aspiration; prenatal testosterone exposure; sound change from below; gender/sex differences; Tyneside English

## 1. Introduction

When discussing sound change, Labov (1966 [2006]: 209) distinguishes sound change from above and sound change from below awareness, or consciousness.

*Editor:* Kleanthes K. Grohmann, University of Cyprus, Cyprus

*Received:* 25 November 2019

*Accepted:* 16 May 2020

*Published:* 8 June 2020

ISSN 1450–3417



CC BY 4.0 License  
© 2020 The authors

Sound change from above refers to the situation in which speakers associate an innovative variant within their community with prestige and in which explicit comments on such variants by the speakers of the relevant community can be found (see e.g., Labov 1966, who reports that higher rates of postvocalic /r/ correlate with higher prestige in New York City English). Sound change from below then refers to those sound changes where the adoption of the innovative feature is not associated with prestige and where the speakers of the community do not comment on the innovation explicitly (see e.g., Baker et al. 2011 for s-retraction, whereby /s/ in words such as *street* sounds more like a /ʃ/; and Lawson et al. 2008, who report de-rhoticisation of postvocalic /r/ in Scottish English as change from below).<sup>1</sup> It is widely reported in quantitative sociolinguistics that there are measurable differences in the frequencies at which speakers use linguistic variables cross-linguistically (e.g., Labov 2001). This is particularly true of ongoing sound changes below the level of consciousness (as defined in Labov 1966 [2006]: 209), where female speakers have been found to “lead” (e.g., Labov 1990, 2001). That is, controlling for other social variables, female speakers tend to use variants that are more progressive within their communities (for continuous sound changes, such as vowel fronting—e.g., Labov, Rosenfelder and Fruehwald 2013, Przedlacka 2001: 47, Hinrichs, Bohmann and Gorman 2013), and also tend to use the new variant more frequently than male counterparts (for categorical variants, such as the presence of post-vocalic /r/ in New York City English—Labov 1966, 1994, 2001). In other words, during the spread of a change from below through a community, the intraspeaker variation of female speakers tends to be a bit different than that of male speakers: on average, a female speaker can be expected to produce more advanced variants more often than a male speaker within the same generation.

This paper focuses on sound change from below and the mechanisms whereby women often end up leading in this type of change. We may wonder what might underlie the observation that women tend to learn or adopt new variants at a higher rate than men. Two major suggestions have been put forward. On the one hand, Labov (1990: 219) suggests that women are most frequently the primary caregivers in most communities, but that boys learn to distance themselves from female norms during their childhood, while girls model their caregivers more faithfully. On the other hand, Eckert (2011) has suggested that women are more likely to engage in “social engineering” and “symbolic domination” than men, due to the setup of the gender roles in their communities, which makes them more likely to be linguistic innovators.<sup>2</sup> Despite these two major attempts to explain why female speakers lead in sound change from below, there is nevertheless no current consensus as to why this is the case. Furthermore,

---

<sup>1</sup> For the purposes of this paper, we are not concerned with the difference between ‘awareness’ and ‘consciousness’. We suspect that these terms could be used to differentiate different intermediate orders of indexicality, but for our purposes, anything that a speaker or community could explicitly identify and discuss as being variant will count as ‘above’ and anything that goes unremarked upon will count as ‘below’.

<sup>2</sup> There is also the proposal according to which male speakers are also agentive in that they purposely differentiate themselves from female speakers (e.g., Trudgill 1972). However, for our purposes this falls within a generally agentive group of explanations.

proposals predicated on the agentivity of speakers in a “linguistic marketplace” are hard-pressed to explain an effect that specifically applies to changes from below the level of consciousness, which do not confer prestige, and specifically not to changes from above the level of consciousness. In particular, when referring to sound change from below, we are concerned with the concept generally accepted within the field of the theory of sound change and sociolinguistics: namely, the process whereby an innovative variant spreads throughout a community, this community comprising multiple generations. We do not focus on the emergence of innovative variants.

Therefore, this paper proposes an account in which subtle biases in learning and cognition contribute to the observed patterns in community change from below (i.e. across speakers), by influencing how speakers represent the interspeaker variation in their internal systems. Particularly, neurological development (which modulates learning) is affected by exposure to prenatal testosterone (see review in Balthazart 2011), which is sexually dimorphic. We suggest that it is a small effect of prenatal T on learning that accounts for the reported differences between men and women, rather than necessarily their active construction of different social roles. Our proposed model is compatible with the neuroendocrinological literature and the sociolinguistic literature, and may make substantial contributions to our understanding of change from below.

Models including any role for neurological/biological factors have rarely been considered seriously (e.g., Labov 2001: 291), and when they have been, they have run the risk of gender essentialism. One of the reasons behind this avoidance of biological factors is due to the methodological challenges of engaging with what these factors entail. For instance, regarding research in biological ageing, establishing the appropriate methods is relatively recent (e.g., Belsky et al. 2015) and many methods available are invasive or otherwise prohibitive. Therefore, these methods are less likely to be accessible to linguists for reasons including ethics of invasive or medical procedures, financial limitations, and importantly the relevant interdisciplinary expertise. However, there are also a number of methods that are non-invasive and financially viable (e.g., Belsky et al. 2015, Wong and Hines 2016). Although engaging even with non-invasive methods is still a cross-disciplinary challenge, this is not insurmountable, as we will show in this paper.

We propose a potential mechanism whereby female speakers typically end up as leaders in sound change from below in the absence of prestige factors as potential motivators. We suggest that one source of this difference could be individual differences in learning from human models, modulated by biological factors. More specifically, it may be the case that the way the prenatal hormone milieu differentially influences development of the brain, the organizing effects of sex hormones on the brain, creates underlying biases in social learning that are different from speaker to speaker, and broadly different between the sexes. These individual biases in social attentiveness then result in individual differences in how speakers sample the linguistic variation they observe in their community and form an internal model for that variation in their linguistic competence (see e.g., Yang 2000, 2002, for a more full discussion of the acquisition of intraspeaker variation as a function of sampling community variation). While this could

broadly result in the differentiated use of linguistic variants by men and women during change in progress, we suggest that the male/female effect is derivative from a gradient biological property, and the underlying effect is driven by a continuous bias regarding attentiveness to social models in one's age cohort. This proposal is based on two types of evidence: research from other fields, including neuroendocrinology and developmental psychology, and an exploratory case study of our own. By engaging with the methodology used in neuroendocrinology and developmental psychology, our study first shows that it is possible to extend the methodology to the field of linguistics, and also presents important results suggesting that the organizing effects attributed to prenatal exposure to sex hormones lead to a small but consistent bias in perception and learning that underlies some apparently sex-dimorphic language use. We suggest that the continuous biological factor of prenatal exposure to androgens has a gradient effect on the distribution of a linguistic variant. Our case study looks into pre-aspiration before voiceless obstruents in Tyneside English (see Section 2.2), a phenomenon which has been reported to be on the rise in this variety of English, with Watt & Allen (2003) showing that younger individuals produce pre-aspiration more frequently. We test this hypothesis by investigating individual differences in the duration of pre-aspiration within a single sex cohort. We use the ratio of the length of the index finger to the ring finger as a continuous biological proxy for early life androgen exposure (2D:4D, see section 2.1 for more details), a measure which has been correlationally linked to the testosterone/estradiol ratio in amniotic fluid in humans (Lutchmaya et al. 2004), and experimentally linked to testosterone exposure in rats (Talarovičová et al. 2009). We propose a model, supported by the empirical observations we report here, in which social learning interacts with biological factors, together contributing to observed patterns of variation during an unconscious sound change in progress (i.e. "change from below").

The field of behavioral neuroendocrinology has shown that the prenatal hormonal milieu of developing organisms bears a potentially causal relation to various adult behaviors in both humans and non-humans (Hines et al. 2004, Hines 2006, Balthazart & Adkins-Regan 2002, Balthazart 2011). Prenatal exposure to testosterone in humans influences various childhood behaviors, including the later adoption of stereotypically gendered behaviors (sometimes referred to as "sex roles", "gender roles", or "gender" more broadly, e.g., Ackerman 2019): For prenatal T on childhood behavior see Pasterski et al. (2005), Auyeung et al. (2009), Cohen-Bendahan et al. (2005), and Hines et al. (2016); for social learning of gender-related labels for objects as in toys that are "for girls" or "for boys" see Hines et al. (2016); for infant eye-contact Lutchmaya et al. (2002); and for lateralisation of certain cognitive functions Liu et al. (2017) and Manson (2008). We would suggest that what unites these various behavioral results is a relationship between prenatal T exposure and a sensitivity to the social models an individual learns from. It could be that higher pre-natal testosterone exposure (T) leads to less sensitive social learning, with learners sampling a wide variety of behaviors they see around them, whereas lower pre-natal T leads to greater pickiness on the part of learners, causing them to focus on people closest to them in various characteristics (e.g., sex, age, etc). Thus, we propose that these behaviors indicate an underlying

mechanism that could also influence the pattern of uptake and spread of new linguistic variants below the level of awareness, apparently led by women due not to purely social factors but also neuroendocrinological organizational effects. We predict that proxies for prenatal testosterone exposure levels, here the 2D:4D digit ratio, will have a significant effect on the amount speakers use a more innovative variant.

Sections 2 and 3 present our exploratory experiment that looks into whether prenatal testosterone exposure correlates with an innovative variant within a single-sex cohort. We then outline how a biosocial model may work in detail in Section 4, where we also flag areas of this proposal that require further investigation.

## 2. Materials and Methods

### 2.1. *Quantifying Prenatal Androgen Level*

While it is difficult to access information about the early-life hormonal milieu in non-clinical human populations, indirect biomarkers do exist. The ratio of the second digit (index finger) to the fourth digit (ring finger), herein referred to as 2D:4D, has been shown in infants to reflect their mothers' amniotic ratios of testosterone to estradiol (Lutchmaya et al. 2004), which makes it a useful albeit noisy proxy for retrospectively measuring prenatal androgen exposure in humans (Cohen-Bendahan et al. 2005, Wong and Hines 2016). Though this biomarker cannot be ethically validated in humans experimentally, it has been shown that experimental manipulation of prenatal testosterone in rat fetuses results in a smaller 2D:4D claw-length ratio after the rats are born (Talarovičová et al. 2009). The effect is more pronounced on the right hand for both rodents (Talarovičová et al. 2009) and humans (Cohen-Bendahan et al. 2005), and right-hand 2D:4D has been shown to predict a level of conformity to aspects of stereotypical gender roles within a single-sex cohort (Atkinson et al. 2017, Brown et al. 2002).<sup>3</sup>

In order to calculate individuals' 2D:4D ratio in our study, participants were asked to place their right hand palm-down on a Doxie portable scanner, producing a portable networks graphic (PNG) image at 600 dpi resolution. The length of the second and fourth digits were then measured from the basal crease to the tip of the finger with digital calipers available in GNU Image Manipulation Program (GIMP). This procedure was chosen because digital measurement of scanned images has been shown to be the most accurate and consistent method of calculating 2D:4D ratio (Allaway et al. 2009).

### 2.2. *Identifying and Quantifying Pre-aspiration*

Pre-aspiration is defined here as a period of voiceless, primarily glottal friction which occurs in the sequences of sonorants and phonetically voiceless obstruents, represented as a superscript *h* in words such as *mass* [ma<sup>h</sup>s] and *mat* [ma<sup>h</sup>t<sup>s</sup>] (Nance

---

<sup>3</sup> This is far from a claim that the sex roles or gendered behaviors themselves are innate. Rather, it is only an observation that this measure bears a relationship to conformity.

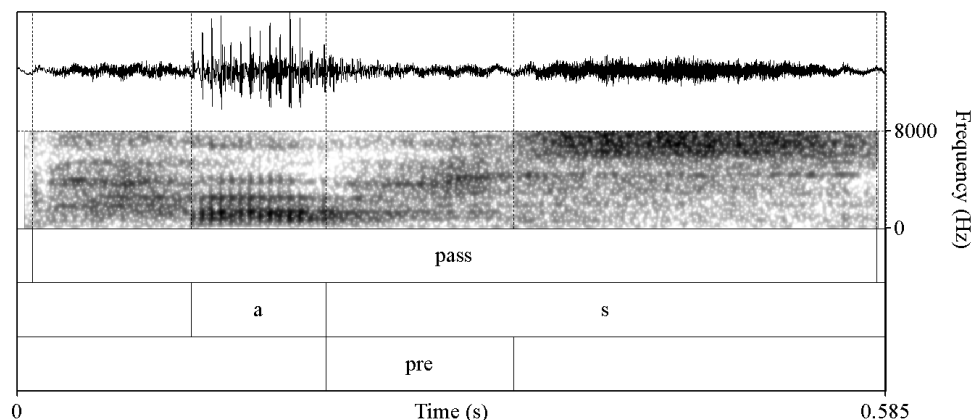


Figure 1: Identification of pre-aspiration in a fricative context.

and Stuart-Smith 2013, Hejná and Scanlon 2015). We identified the presence of pre-aspiration and quantified its duration as described below (see Hejná 2016 for more details).

In the plosive environment (e.g., *mat*), the left boundary of pre-aspiration was determined by the absence of voicing associated with the vowel and the right boundary by the absence of pre-aspiration friction. In the fricative environment (e.g., *mass*), the identification of the left boundary was the same as in the plosive context. The right boundary, however, could not be determined on the basis of the lack of friction because fricatives are by definition articulated as turbulent sounds that lack closure, just like pre-aspiration. We therefore used more detailed spectral criteria for the right boundary. While /f/ and /s/ show a concentration of high-intensity energy at higher frequencies, glottal friction shows low-intensity energy spread across the frequencies, as shown in Figure 1. The highlighted portion in the figure delimits pre-aspiration in the word *pass*. The voicing associated with the vowel stops at the left-hand edge of pre-aspiration, and the high-frequency band of energy associated with the friction of /s/ begins at the right-hand edge of pre-aspiration.

In the cases of /θ/ and /f/, the criterion based on the distinct centers of energies is not always sufficiently reliable because the two oral fricatives are associated with a fairly even spread of energies across the frequencies as well, in contrast to /s/ and /ʃ/ (Fry 1982). In these cases, it was the structure of the bands of higher energy in the pre-aspiration friction (i.e., formant structure) that distinguished it from the friction of the two oral fricatives: pre-aspiration formant structure tends to closely mirror that of the local vowel, which is distinct from the typical patterns of energy distribution in oral fricatives.

Only positive cases of pre-aspiration (i.e., durations greater than 0 ms) were included in the analyses. Pre-aspiration duration was normalized as a natural logarithm of the percentage of the overall word duration in order to account for potential speaking rate differences and prosodic lengthening effects.

### 2.2.1. Descriptive Statistics

The data comprise 1,533 tokens with pre-aspiration, ranging from 13 to 157 tokens per speaker, with an average of 69.7 tokens and a median of 55 tokens. The words analyzed contain the following structures: CVC(C)(C) (as in *pass*, *past*, *lasts*) and CVC(C)V (as in *messy*, *misty*); in which the first syllable is always stressed and the second syllable is not if the word has more than a single syllable. Our analysis only includes pre-aspiration in lexically stressed contexts, since this is where pre-aspiration has been reported to be produced most frequently (e.g., Hejná 2015: Chap. 3, and the references therein). With these structures, we coded for the variables of vowel quality (in the stressed syllable; 15 levels, see Figure 2) and the immediately adjacent, pre-aspiration inducing obstruent (eight levels, see Figure 3).

These phonetic environments of the tokens were not balanced across speakers because this cannot be controlled for in naturalistic interview data intended to produce casual speech. Thus, the preceding vowel phoneme and the place of articulation of the local consonant are included as factors in the statistical analyses.

In addition, we also coded for the linguistic context immediately following the word analysed; namely, we distinguished following words as (1) those starting with a vowel (e.g., ...*cat and dog*); (2) those that started with a consonant (e.g., ...*cat running*); and (3) pauses (e.g., ...*cat.*). This was done because in phenomena such as /t/-flapping, flapping applies in sequences such as *get it* and *ate it*. Since we did not know based on the literature available whether this may also affect pre-aspiration, it was deemed prudent to code for this potential factor. Finally, the

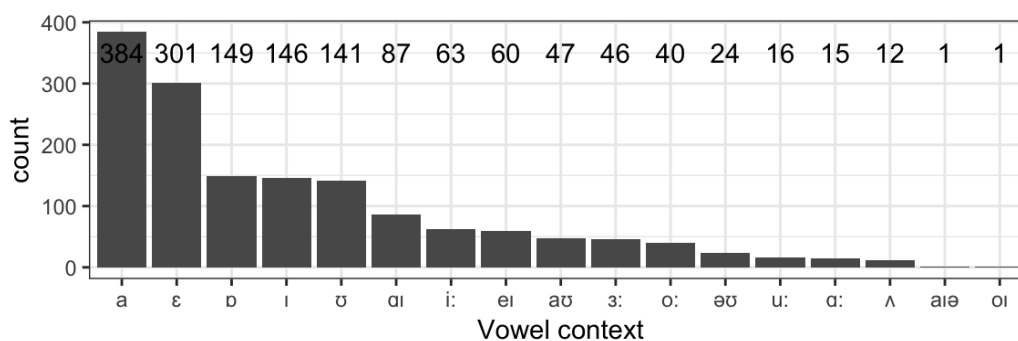


Figure 2: Distribution of vowel phonemes preceding pre-aspiration in our dataset.

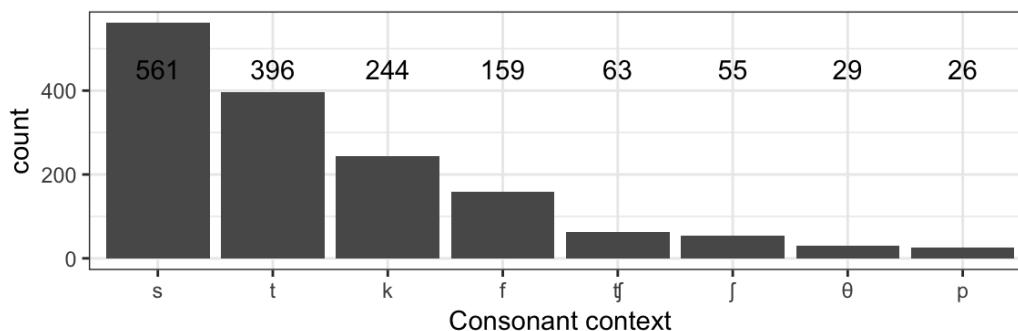


Figure 3: Distribution of obstruent phonemes inducing pre-aspiration in our dataset.



pause context was included because of domain final lengthening effects as well. Tokens with pre-aspiration were followed by a consonant in 506 cases, by a pause or silence in 598 cases, and by a vowel in 429 cases.

### 2.3. *Participants*

Twenty-two participants identifying themselves to researchers as women/female were recruited from the Tyneside area and interviewed using standard sociolinguistic procedures (Tagliamonte 2006). They ranged in age from 22 to 45 years of age, with a mean age of 30.9 (SD = 8.7) and median age of 29.5. Each person was asked a series of prompts in order to elicit a naturalistic speech sample, which averaged approximately one hour in duration. All participants gave signed consent, and were provided with information about the study, investigators, and university ethics procedures. Twenty of the participants were White, of European heritage. One participant was primarily of East Asian heritage, and one participant was primarily of South Asian heritage. As we did not have enough participants from different ethnicities to reliably model any potential differences by ethnicity (Manning et al. 2007), we did not include ethnicity as a separate variable in the analytical models.

### 2.4. *Recording and Data Processing*

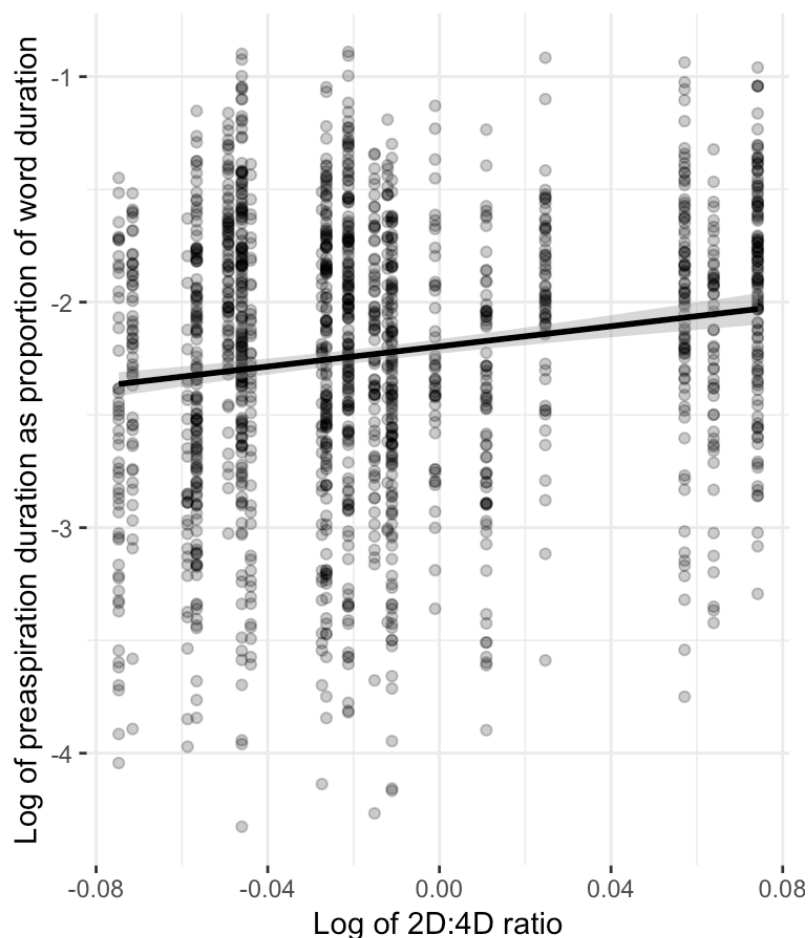
Interviews were recorded one-to-one on a Zoom H4n Handy Recorder in a quiet room, using the built-in microphone. The data were sampled at 44.1 kHz and stored in .wav format. Interview transcriptions were force-aligned to the recorded sound files using FAVE-align (Rosenfelder et al. 2014), then manually checked for accuracy in the relevant words. The phonetic segments preceding and following the pre-aspiration were extracted, along with the lexical environment and syllabic stress. Pre-aspiration was annotated manually. Ambiguous cases were excluded from the analyses. Datasets used for the analyses above are publicly available (see Data Availability statement).

## 3. **Results**

To assess the contribution of prenatal exposure to androgens to the usage of pre-aspiration, we compared a series of linear mixed effects regression models based on the principle of parsimonious model design (Bates et al. 2018). The contribution of the 2D:4D ratio was the independent variable of interest ('logDigit'), and the variation in pre-aspiration duration as a proportion of word duration ('logProp') was the dependent variable. Other fixed effects were speaker age, the interviewer (two levels), the vowel context (see *Figure 2*), the consonantal segment triggering pre-aspiration (see *Figure 3*), the metrical foot position (medial, as in *better*, or final, as in *bet*), and the following segment (consonant, vowel, or pause). A random intercept for speaker was included but the structure of random slopes was limited by the naturalistic nature of the data. Since the factors that are logically able to vary randomly by speaker contain a large number of categorical levels (e.g., triggering consonant environment, vowel context), these factors were too

imbalanced to include in the analysis as interactions. This analysis was performed using the lme4 package (Bates et al. 2015) in R (R Core Team) using the following formula: `lmer(logProp ~ logDigit + age + interviewer + vowelContext + triggerConsonant + footPosition + followingSeg + (1 | speaker), data=data, REML=FALSE)`. To calculate the contribution of the relevant factor to the overall model fit, a chi-square test was used to compare this full model to a nested subset model which had the single relevant term removed.

The linear mixed effects regression analysis revealed a significant main effect of 2D:4D digit ratio, with higher ratios (corresponding to lower androgen exposure) correlated with longer proportional pre-aspiration durations ( $\beta = 1.31$ ,  $SE = .47$ ,  $\chi^2(1) = 6.6$ ,  $p = 0.010$ ). This small but significant effect is confirmed by a Pearson's product-moment correlation test ( $t = 6.2$ ,  $df = 1531$ ,  $p < 0.0001$ ; CI 95% = 0.11–0.20;  $r = 0.16$ ). These results suggest that speakers who were exposed to lower levels of androgens during prenatal development also produce longer pre-aspiration in Tyneside English, as illustrated in *Figure 4*. What *Figure 4* shows is that the effect of 2D:4D ratio on the log of proportional pre-aspiration durations is small; yet, it is nevertheless statistically significant. This is consistent with the hypothesis that prenatal hormone organizing effects could contribute to the



*Figure 4: Lower prenatal androgens correlate with more advanced variant.*

observed differences in language variation from below the level of consciousness, which have typically been attributed to social effects of gender.

Speaker age did not significantly contribute to this model ( $\beta < 0.01$ ,  $SE < 0.01$ ,  $\chi^2(1) = 2.1$ ,  $p = 0.14$ ). Age was expected to contribute to the variation of a sound change in progress, since such changes typically are detectable in apparent time; that is, people of different ages demonstrate different stages of the sound change when contemporaneous speech samples are compared. It is possible that the sound change in question had been completed at the time of interviews, thus no change in apparent time is detectable. It is also possible that our sample of speakers is (relatively) homogeneous and uniform in age, leading to a lack of power in this particular dimension. However, the significant correlation between 2D:4D ratio and pre-aspiration suggests that this line of research may prove fruitful in future investigations.<sup>4</sup> We now turn to discussing a biosocial mechanism whereby sound changes from below often emerge as led by female speakers in their communities.

## 4. Discussion

### 4.1. A biosocial model

Our study provided evidence that, even within a single-sex sample, prenatal testosterone exposure has an inverse association with the duration of pre-aspiration. In other words, lower levels of prenatal testosterone correlate with longer pre-aspiration duration. Since this parallels observations made in the wider literature which are attributed to sex or gender differences, we suggest that the often observed sex effect may actually stem from a more basic effect, as sex cohorts will typically also differ in their hormone profiles. We propose a possible mechanism that can account for the previously observed sex differences in studies of sound change in light of the observed individual inter-speaker variation. The suggestion is to some extent in line with Labov (1990), who proposes that the explanation may lie in the learning conditions of future language users, who tend to be exposed more frequently to females due to their primary caregivers being female. However, while we agree that learning mechanisms may be key, our proposal diverges from Labov's in several ways. Firstly, we propose that the learning strategies involved vary not only across but also within sex cohorts. Secondly, this can be independent of whether the primary caregiver is female or male; rather, differences in learning strategies can be related to variation in prenatal testosterone exposure. Regarding Eckert's (2011) line of explanation, ours differs from hers in that we exclude agency on the part of the speaker-adopter as the sole driving mechanism. Crucially, however, the model we propose is assumed to co-exist with possible agentive factors discussed by Eckert, linked to group membership.

Uncontroversially, individuals differentially learn and reproduce linguistic features undergoing population-level change, leading to subtle inter-speaker

---

<sup>4</sup> The variable of interviewer (with two levels: one female from Newcastle, one male from San Francisco) did not show any effects on pre-aspiration duration.

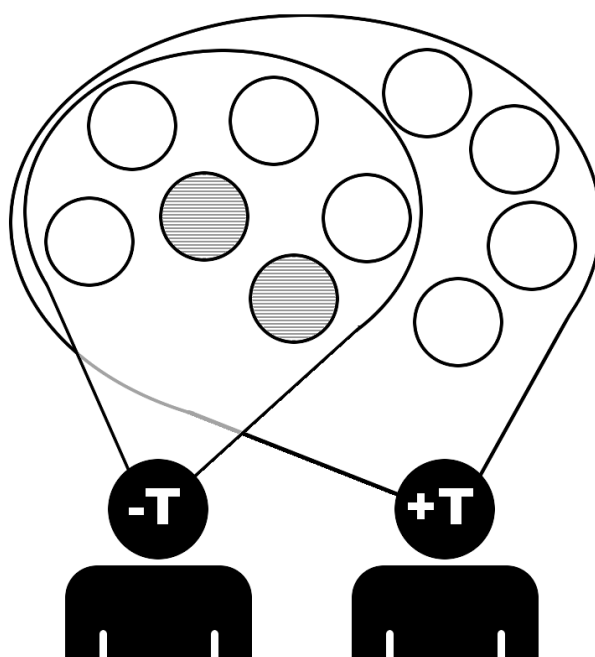


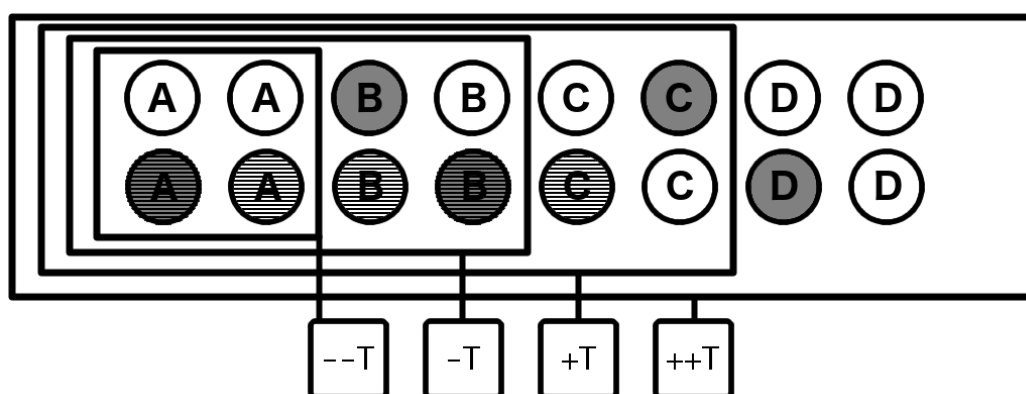
Figure 5: A biosocial model of attention during language acquisition.

individual differences in linguistic behavior. Building on the research related to social learning patterns and endocrine organizing effects we mentioned above in Introduction, we posit a biosocial model that takes seriously these influences on early-life development, and suggests a mechanism for the sexually dimorphic behavior during language change that is missing from the extant literature (e.g., Labov 1990, Eckert 2011). We hypothesise that the differences in infant eye-contact Lutchmaya et al (2002) report between higher prenatal T infants and lower prenatal T infants is an indication of generally better social attentiveness in lower prenatal T individuals, throughout life. If that is the case, then individuals exposed to higher levels of prenatal testosterone (+T) would be less attentive to appropriate adult social models than individuals exposed to lower levels of prenatal testosterone (-T). We refer to this variation in attentiveness as “attentional specificity”, which means being sensitive to all potentially socially salient characteristics of people in the population, including their ages.

Figure 5 illustrates such an effect of attentional specificity on sound change in the following way: In this illustration, circles correspond to individuals in a population of speakers. Shaded circles indicate speakers with a high frequency of a new linguistic variant or a more advanced versions of a continuous variant: suppose the actuation of some new change from below has already occurred, and a new variant has begun its spread. The left-hand individual (-T, lower testosterone exposure) attends more to social models who are more similar in terms of social factors such as age, thus samples the speech of only a subset of the population which they are exposed to. In this subset, the new linguistic variant is heard in one third of the observations. The right-hand individual (+T, higher testosterone exposure) is less selective in their attentiveness to social models, and so samples randomly from the entire population. This speaker hears the new

variant proportionally less often because only one fifth of their sample observations contain it. If speakers store the frequencies of variants they hear in language acquisition and use these to form an internal model of the variation and thus their target frequencies of the variants for production (as extensively argued in Yang 2000, 2002, among others), the  $-T$  speaker will then produce a higher frequency of the new linguistic variant than the  $+T$  speaker. Thus, individual variation in social attention, as biased by prenatal hormone organizing effects, influences the proportion of tokens of a new variant that an individual samples during language acquisition and learning. Because attentional specificity influences who in the population an individual attends to and learns from, individuals with lower prenatal exposure to testosterone are more likely to learn from those in their peer group (i.e., with whom they are closest in age). As children learn their first language, these individuals will attend to a younger subset of the speaker population to which they are exposed. One can think of this as a model of how learners sample from all the community productions they hear, and this sample is a given learner's primary linguistic data (see the example sampling scenarios in *Figure 5* and *Figure 6*).

In *Figure 6*, this process is illustrated in more detail than in *Figure 5* by four variants of an individual that differ in prenatal exposure to testosterone, ranging from least ( $--T$ ) to most ( $++T$ ); note that these categories are for illustrative purposes, and amount of pre-natal T exposure is, in reality, a continuous variable. In individuals with typical hormonal profiles, we may assume females with lower prenatal testosterone exposure correspond to  $--T$ , females with somewhat higher prenatal testosterone exposure correspond to  $-T$ , males with somewhat lower prenatal testosterone exposure correspond to  $+T$ , and males with higher prenatal testosterone exposure correspond to  $++T$ . All individuals are within the same age cohort (Cohort A), with Cohort B being slightly older than A, Cohort C older than B, and Cohort D representing the rest of the population at large. Circles filled in with grey backgrounds indicate speakers producing a variant that is spread mostly evenly across the population. Circles with horizontal stripes represent speakers producing a variant that has been innovated in younger cohorts, thus is concentrated in younger cohorts at this point in time. *Table 1* describes the differ-



*Figure 6: Example of how this model would play out in a (simplistic but illustrative) population. An interactive version of this model is available at: <https://lmackerman.shinyapps.io/aplymodel-201904> (as of 6 May 2019).*

Individual	Grey circles (evenly spread throughout population)		Striped circles (concentrated in younger cohorts)	
	Number of social models producing new variant	Proportion of social models producing new variant	Number of social models producing new variant	Proportion of social models producing new variant
++T	5	31.3 %	5	31.3 %
+T	4	33.3 %	5	41.7 %
-T	3	37.5 %	4	50 %
--T	1	25 %	2	50 %

Table 1: Rate of exposure based on individual variation in attention to social models, i.e., differential attention to speech produced by similar age-cohort models modulated by prenatal testosterone exposure.

ential rates at which each of the four listeners would be exposed to the two types of features, as relevant for the example shown in Figure 6.

From this illustration (Figure 6), individuals are exposed to the “grey” feature at similar rates, with slight variation between the --T and -T individuals due to a quirk of the distribution in the population. If we assume that the --T and -T individuals are both female and the +T and ++T individuals are male, this pattern of exposure would result in little to no variation in uptake across the population. In contrast, the “striped” feature that skews young is adopted at higher rates by the female individuals than the male individuals; the new variant spreads over time in both men and women, but females will be consistently ahead of males at every time point in the spread. Given a binary categorization by sex, this could appear to be a property that varies categorically. However, when individuals are assessed on a gradient scale, we can see that this pattern is predicted by prenatal exposure to testosterone and the individuals’ social network in tandem.<sup>5</sup>

The more selectively attentive individuals, as they acquire language in early life, sample speech from a subset of the population that produces more of the new variant which is concentrated in younger speakers (who match in age with the learners). If less exposure to prenatal testosterone is associated with more attentional specificity, female speakers will, on average, show larger amounts of new linguistic variants undergoing change below awareness than male speakers of the same age throughout the inter-generational spread of a new linguistic variant.

Crucially, however, there are four important things that need to be added. First, because the distribution of an innovative variant is not uniform across a speech community, nor is—presumably—that of individuals with specific levels of prenatal testosterone exposure, it cannot always be the case that the most innovative variants will be found with highest rate of application / most advanced realization in the following order: females with relatively lower testosterone exposure (--T) > females with relatively higher testosterone exposure (-T) > males with relatively lower testosterone exposure (+T) > males with relatively higher testosterone exposure (++T). Second, changes from below show clear apparent time distributions: we are more likely to find innovations in

<sup>5</sup> Here we have only focused on age-related aspects of social networks, and in a rather simplistic way in comparison to highly complex real-life situations. The proposed model needs to be developed to deal with social networks in a more in-depth ways in future.

younger age cohorts (e.g., Tagliamonte 2016). This fact alone makes it more likely that younger cohorts are exposed to those speakers with the innovative features in age stratified cultures.

Thirdly, the mechanism we propose is additional to the bias for younger speakers to be the source of innovative features. In other words, the prenatal testosterone exposure bias adds to that of innovative features typically arising in children and peaking in adolescent cohorts. These individuals with higher attentional specificity will then learn and produce more of the new variant than those in their age cohort with lower attentional specificity, since these latter speakers will have sampled more broadly from the relevant population. As the proportion of the population using the new variant increases, those with lower attentional specificity will become more likely to pick it up, but they will do so at a lower proportion than the high attentional specificity speakers. This lag is due to dilution of the new variant in a differentially larger population sample for individuals with lower attentional specificity. Note that the mechanism we propose here for the sex-differential lag is independent of the mechanism that is responsible for the incrementation of the change from below. It is, for instance, entirely compatible with a “momentum-based” model for incrementation (Bermúdez-Otero, to appear). In that scenario, adolescents increment the frequency or advancedness of a new variant in line with their representation of an age vector for that variant in the population. Our mechanism would simply mean that the represented age vector is systematically a little different in  $-T$  vs  $+T$  individuals, because they have estimated it on the basis of a slightly different sampling of the population. Their incrementation based on this age vector will then also have slightly different outcomes, even if  $-T$  and  $+T$  speakers increment by the same amount.

In what follows, we also discuss two important methodological factors that may have bearing on the model we suggest and which need to be taken into account in future research.

## 4.2. *Further considerations*

### 4.2.1. *Laryngeal physiology*

One potential confound in assessing the plausibility of the attentional specificity mechanism is the possibility that prenatal testosterone exposure could have a direct effect on the physiology of the vocal apparatus, thus producing the observed pre-aspiration pattern. In other words, we must address the concern that the effect reported may be attributable to laryngeal structure differences rather than those related to brain structures involved in learning and attention. This is especially the case since the larynx is generally seen as a secondary sexual organ (e.g., Abitbol et al. 1989, Amir & Biron-Shental 2004, Collins & Missing 2003, Hall 1995, Henton & Bladon 1985), with its sexual characteristics resulting from circulating (activational) hormone effects (see e.g., Raj et al. 2008, Wadnerkar et al. 2006, Whiteside et al. 2004, for the effects of menstrual cycle on voice quality and laryngeal properties of consonants), although establishing whether and where there are hormonal receptors in the laryngeal structures is controversial

(Nacci et al. 2011). Interestingly, Fouquet et al. (2016), who conducted a real-time study of  $f_0$  of 10 male speakers at 7-year intervals starting at 7 and ending at 56 years of age, found that intra-sex differences are apparent already at the age of 7 and persist to adulthood. The authors suggest this may be due to prenatal androgen exposure differences (2016: 7) rather than those of circulating hormones that lead to vocal differences across the two sexes during puberty; however, information related to prenatal testosterone exposure was not available to the research team. Ferdezi et al. (2011) have presented a study which looked into whether there is a correlation between their male speaker's  $f_0$  and their 2D:4D of either the left or the right hand: no correlation was found, suggesting that the laryngeal structures are not affected by prenatal testosterone exposure.

Additionally, it is indicative that the effect of 2D:4D holds for pre-aspiration, but not for the other laryngeal phenomena that we quantified in addition to pre-aspiration, namely local breathiness and creaky voice (see e.g., Hejná 2015: Chap. 3, for local breathiness, and Keating, Garellek & Kreiman 2015 for creaky voice, respectively). As shown in *Figure 7*, while the correlation between 2D:4D and pre-aspiration duration is significant, no such correlations are observed for breathiness or creakiness in our data.

This suggests that the effect is not due to prenatal testosterone exposure on the laryngeal structures. On the other hand, different laryngeal phenomena rely on different laryngeal gestures and it is not known which laryngeal structures should be targeted by prenatal testosterone exposure. For this reason, future work should include a range of linguistic phenomena that rely on different parts of the vocal tract to conclusively rule out effects of prenatal testosterone exposure on the laryngeal structures rather than on general learning mechanisms associated with brain structure differences affecting linguistic production.

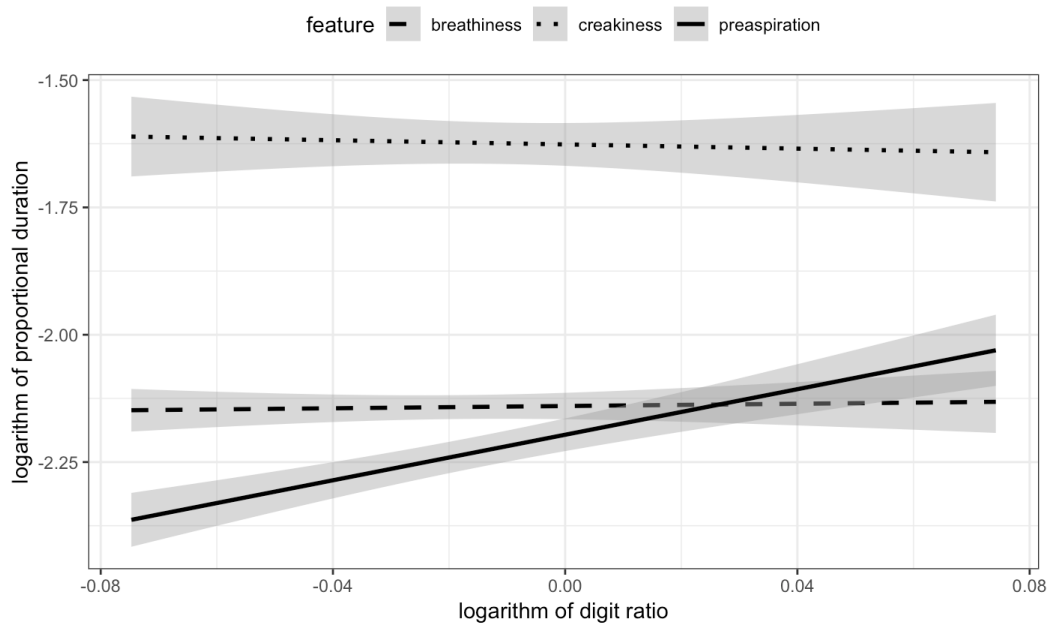


Figure 7: No correlation between 2D:4D and breathiness and creaky voice. Breathiness:  $r = 0.0078$ ; creakiness:  $r = -0.0150$ , pre-aspiration:  $r = 0.1555$ .



#### 4.2.2. 2D:4D

Another caveat of this study is the use of the 2D:4D ratio, which is controversial in its use as a biomarker for human exposure to prenatal testosterone. The previous studies on the relationship between testosterone and 2D:4D in humans and rats, respectively, do not examine precisely the same biological relationships, thus must be regarded carefully (Lutchmaya et al. 2004, Talarovičová et al. 2009). Additionally, there are other differences in the timing of sexual organizing effects in rats and humans, and the way sex hormones enter the brain may be different in the two species (Balthazart 2011). See also reviews of 2D:4D analyses, which note various inconsistencies in reported studies linking 2D:4D and behavior (Cohen-Bendahan et al. 2005, Wong & Hines 2016). Future work might circumvent these issues by using other biomarker proxies, potentially including spontaneous or click-evoked otoacoustic emissions (McFadden 2002, McFadden & Pasanen 1998, 1999), or ideally, by accessing longitudinal data gathered from a population whose prenatal testosterone levels were measured.

### 5. Conclusion

In this study, we demonstrate that a potential biomarker for prenatal androgen exposure (2D:4D) predicts the duration of pre-aspiration, a sound change in progress in Tyneside English (albeit with a weak correlation). Since prenatal testosterone has also been shown to co-vary with some aspects of later social behaviors, we suggest a mechanism whereby social attentiveness relates to the propagation of linguistic variants over a number of generations. In particular, since lower levels of prenatal testosterone predict more advanced linguistic variants during a change, we suggest that people who were prenatally exposed to lower levels of testosterone will be more likely to learn new linguistic variants during a sound change in progress, due to a small bias in the social models they attend to, although future modelling should confirm how frequently this arises in a range of realistic social network setups vis-à-vis linguistic innovations. The small bias in attending to social models like themselves affects the way different individuals sample the population of speakers they are exposed to.

While these results are not yet fully conclusive, they do indicate that biological factors may have a role in explaining some patterns of language change observed throughout the sociolinguistic literature. For changes below the level of awareness, this small but persistent influence of prenatal testosterone organizing effects could explain the apparently sex-correlated observations in the literature. We hope that the suggestive results reported here provide a starting point for further investigation into how small, individual differences in biological development influence social attention, learning styles, and linguistic learning.

### Data Availability

The data used for analysis in this article are freely available from the Open Science Framework: <http://doi.org/10.17605/OSF.IO/P5NKA>

## Acknowledgements

We would like to thank Claire Cochrane for her immense help with data collection and participant recruitment. We are also grateful to the Centre for Behaviour and Evolution seminar audience at Newcastle University. This work was, in part, supported by a Wellcome Trust Capital Award to Newcastle University (grant number 092504). All errors are, of course, our own.

## References

- Ackerman, Lauren. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: A Journal of General Linguistics* 4(1), 117. [doi:10.5334/gjgl.721](https://doi.org/10.5334/gjgl.721)
- Allaway, Heather C., Terri G. Bloski, Roger A. Pierson & Marla E. Lujan. 2009. Digit ratios (2D:4D) determined by computer-assisted analysis are more reliable than those using physical measurements, photocopies, and printed scans. *American Journal of Human Biology* 21(3), 365–370.
- Abitbol, Jean, Jean de Brux, Ginette Millot, Marie-Françoise Masson, Odile Languille Mimoun, Helene Pau & Beatrice Abitbol. 1989. Does a hormonal vocal cord cycle exist in women? Study of Vocal Premenstrual Syndrome in voice performers by videostroboscopy-glottography and cytology on 38 women. *Journal of Voice* 3(2), 157–162.
- Amir, Ofer & Tal Biron-Shental. 2004. The impact of hormonal fluctuations on female vocal folds. *Current Opinion in Otolaryngology and Head and Neck Surgery* 12(3), 180–184.
- Atkinson, Beth M., Tom V. Smulders & Joel C. Wallenberg. 2017. An endocrine basis for tomboy identity: The second-to-fourth digit ratio (2D:4D) in “tomboys”. *Psychoneuroendocrinology* 79, 9–12.
- Auyeung, Bonnie, Simon Baron-Cohen, Emma Ashwin, Rebecca Knickmeyer, Kevin Taylor, Gerald Hackett & Melissa Hines. 2009. Fetal testosterone predicts sexually differentiated childhood behavior in girls and in boys. *Psychological Science* 20(2), 144–148.
- Baker, Adam, Diana Archangeli & Jeff Mielke. 2011. Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change* 23(3), 347–374.
- Balthazart, Jacques. 2011. *The Biology of Homosexuality*. Oxford: Oxford University Press.
- Balthazart, Jacques & Elizabeth Adkins-Regan. 2002. Sexual differentiation of brain and behavior in birds. *Hormones, Brain and Behavior* 4, 223–301.
- Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & Harald Baayen. 2018. Parsimonious mixed models. [arXiv:1506.04967](https://arxiv.org/abs/1506.04967).
- Bates, Douglas, Martin Mächler, Ben B. Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.
- Belsky, Daniel W., Avshalom Caspi, Renate Houts, Harvey J. Cohen, David L. Corcoran, Andrea Danese, HonaLee Harrington, Salomon Israel, Morgan E. Levine, Jonathan D. Schaefer, Karen Sugden, Ben Williams, Anatoli I.

- Yashin, Richie Poulton & Terrie E. Moffitt. 2015. Quantification of biological aging in young adults. *PNAS*, 1–7.
- Bermúdez-Otero, R. To appear. Individual differences and the explanation of sound change. In James Kirby, Lauren Hall-Lew & Patrick Honeybone (eds.), *Individuals, Communities, and Sound Change*. *Glossa special collection*.
- Brown, Windy M., Christopher J. Finn, Bradley M. Cooke & S. Marc Breedlove. 2002. Differences in finger length ratios between self-identified “butch” and “femme” lesbians. *Archives of Sexual Behavior* 31(1), 123–127.
- Cohen-Bendahan, Celina C. C., Cornelieke van de Beek & Sheri A. Berenbaum. 2005. Prenatal sex hormone effects on child and adult sex-typed behavior: methods and findings. *Neuroscience and Biobehavioral Reviews* 29(2), 353–384.
- Collins, Sarah A. & Caroline Missing. 2003. Vocal and visual attractiveness are related in women. *Animal Behaviour* 65, 997–1004.
- Eckert, Penelope. 2011. Language and power in the preadolescent heterosexual market. *American Speech* 86(1), 85–97.
- Ferdezi, Camille, Jean-François Lemaître, Juan David Leongómez & S. Craig Roberts. 2011. Digit ratio (2D:4D) predicts facial, but not voice or body odour, attractiveness in men. *Proceedings of the Royal Society B* 278(1724), 3551–3557.
- Fouquet, Meddy, Katarzyna Pisanski, Nicolas Methevon & David Reby. 2016. Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. *Proceedings of the Royal Society B* 3(10), 160395.
- Fry, Dennis B. 1982. *The Physics of Speech*. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
- Hall, Kira. 1995. Lip service on fantasy lines. In Kira Hall & Mary Bucholtz (eds.), *Gender Articulated: Language and the Socially Constructed Self*, 183–226. New York: Routledge.
- Hall, Kira & Mary Bucholtz (eds.). 1995. *Gender Articulated: Language and the Socially Constructed Self*. New York: Routledge.
- Hejná, Míša. 2015. Pre-aspiration in Welsh English: A case-study of Aberystwyth. Manchester, UK: University of Manchester dissertation.
- Hejná, Míša. 2016. Pre-aspiration: manual on acoustic analyses 1.1. Available from: <https://ling.auf.net/lingbuzz/003184>
- Hejná, Míša & Jane Scanlon. 2015. New laryngeal allophony in Manchester English. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow.
- Henton, C. G. & R. A. W. Bladon. 1985. Breathiness in normal female speech: inefficiency versus desirability. *Language and Communication* 5(3): 221–227.
- Hines, Melissa. 2006. Prenatal testosterone and gender-related behaviour. *European Journal of Endocrinology* 155 (supplement 1), S115–S121.
- Hines, Melissa, Charles Brook & Gerard S. Conway. 2004. Androgen and psychosexual development: core gender identity, sexual orientation, and recalled childhood gender role behavior in women and men with congenital adrenal hyperplasia (CAH). *Journal of Sex Research* 41(1), 75–81.
- Hines, Melissa, Vickie Pasterski, Debra Spencer, Sharon Neufeld, Praveetha Patalay, Peter Hindmarsh, Ieuan A. Hughes & Carlo L. Acerini. 2016.

- Prenatal androgen exposure alters girls' responses to information indicating gender-appropriate behaviour. *Philosophical Transactions of Royal Society B* 371(1688), 20150125.
- Hinrichs, Lars, Alex Bohmann & Kyle Gorman. 2013. Real-time trends in the Texas English vowel system: F2 trajectory in GOOSE as an index of a variety's ongoing delocalization. *Rice Working Papers in Linguistics* 4, 1–12.
- Keating, Patricia, Marc Garellek & Jody Kreiman. 2015. Acoustic properties of different kinds of creaky voice. *Proceedings of 18th International Congress of Phonetic Sciences, Glasgow*.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.
- Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 205–254.
- Labov, William. 1994. *Principles of Linguistic Change. Volume 1: Internal Factors*. Oxford: Blackwell.
- Labov, William. 2001. *Principles of Linguistic Change. Volume 2: Social Factors*. Oxford: Blackwell.
- Labov, William, Ingrid Rosenfelder & Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89(1), 30–60.
- Lawson, Eleanor, Jane Stuart-Smith & James Scobbie. 2008. Articulatory insights into language variation and change: Preliminary findings from an ultrasound study of derhoticization in Scottish English. *University of Pennsylvania Working Papers in Linguistics* 4(2), 102–110.
- Liu, Jinfeng, Dan Wang, Xiaoting Li & Wang Ningyu. 2017. Association between sex and speech auditory brainstem responses in adults, and relationship to sex hormone levels. *Medical Science Monitor* 23, 2275–2283.
- Lutchmaya, Svetlana, Simon Baron-Cohen & Peter Raggatt. 2002. Foetal testosterone and eye contact in 12-month-old human infants. *Infant Behavior and Development* 25(3), 327–335.
- Lutchmaya, Svetlana, Simon Baron-Cohen, Peter Raggatt, Rebecca Knickmeyer & John T. Manning. 2004. 2nd to 4th digit ratios, fetal testosterone and estradiol. *Early Human Development* 77(1), 23–28.
- Manning, John T., Andrew J. G. Churchill & Michael Peters. 2007. The effects of sex, ethnicity, and sexual orientation on self-measured digit ratio (2D: 4D). *Archives of Sexual Behavior* 36(2), 223–233.
- Manson, JoAnn E. 2008. Prenatal exposure to sex steroid hormones and behavioral/cognitive outcomes. *Metabolism* 57, S16-S21.
- McFadden, Dennis. 2002. Masculinization effects in the auditory system. *Archives of Sexual Behavior* 31(1), 99–111.
- McFadden, Dennis & Edward G. Pasanen. 1998. Comparison of the auditory systems of heterosexuals and homosexuals: Click-evoked otoacoustic emissions. *Proceedings of the National Academy of Sciences* 95(5), 2709–2713.
- McFadden, Dennis & Edward G. Pasanen. 1999. Spontaneous otoacoustic emissions in heterosexuals, homosexuals, and bisexuals. *Journal of the Acoustical Society of America* 105(4), 2403–2413.

- Nacci, Andrea, Bruno Fattori, Fabio Basolo, Maria E. Filice, Katia De Jeso, Luca Giovannini, Luca Muscatello, Fabio Matteucci & Francesco Ursino. 2011. Sex hormone receptors in vocal fold tissue: a theory about the influence of sex hormones in the larynx. *Folia Phoniatrica et Logopaedica* 63, 77–82.
- Nance, Claire & Jane Stuart-Smith. 2013. Pre-aspiration and post-aspiration in Scottish Gaelic stop consonants. *Journal of the International Phonetic Association* 43(1), 129–152.
- Pasterski, Vickie L., Mitchell E. Geffner, Caroline Brain, Peter Hindmarsh, Charles Brook & Melissa Hines. 2005. Prenatal hormones and postnatal socialization by parents as determinants of male-typical toy play in girls with congenital adrenal hyperplasia. *Child Development* 76(1), 264–278.
- Przedlacka, Joanna. 2001. Estuary English and RP: some recent findings. *Studia Anglica Posnaniensia* 36, 35–50.
- R Core Team. 2016. R: A Language and Environment for Statistical Computing. Available from: <https://www.r-project.org/>
- Raj, Anoop, Bulbul Gupta, Anindita Chowdhury & Shelly Chadha. 2008. A study of voice changes in various phases of menstrual cycle and in postmenopausal women. *Journal of Voice* 24(3), 363–368.
- Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Suite Version 1.1.3.
- Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
- Tagliamonte, Sali. 2016. *Teen Talk. The Language of Adolescents*. Cambridge: Cambridge University Press.
- Talarovičová, Alžběta, Lucia Kršková & Jana Blažeková. 2009. Testosterone enhancement during pregnancy influences the 2D:4D ratio and open field motor activity of rat siblings in adulthood. *Hormones and Behavior* 55(1), 235–239.
- Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1, 179–195.
- Watt, Dominic & William Allen. 2003. Tyneside English. *Journal of the International Phonetic Association* 33(2), 267–271.
- Wadnerkar, Meghana B., Patricia E. Cowell & Sandra P. Whiteside. 2006. Speech across the menstrual cycle: a replication and extension study. *Neuroscience Letters* 408, 21–24.
- Whiteside, Sandra P., Anna Hanson & Patricia E. Cowell. 2004. Hormones and temporal components of speech: sex differences and effects of menstrual cyclicity on speech. *Neuroscience Letters* 367, 44–47.
- Wong, Wang I. & Melissa Hines. 2016. Interpreting digit ratio (2D:4D)–behavior correlations: 2D:4D sex difference, stability, and behavioral correlates and their replicability in young children. *Hormones and Behavior* 78, 86–94.
- Yang, Charles. 2000. Internal and external forces in language change. *Language Variation and Change* 12(3), 231–250.
- Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

# Children's Learning of a Semantics-Free Artificial Grammar with Center Embedding

Shiro Ojima<sup>1,2</sup> & Kazuo Okanoya<sup>1,\*</sup>

<sup>1</sup> Department of Life Sciences, Graduate School of Arts and Sciences, The University of Tokyo, Japan

<sup>2</sup> College of Education, Yokohama National University, Japan

\* Corresponding author: [cokanoya@mail.ecc.u-tokyo.ac.jp](mailto:cokanoya@mail.ecc.u-tokyo.ac.jp)



SO: <https://orcid.org/0000-0002-2740-692X>

KO: <https://orcid.org/0000-0002-7439-9566>

Whether non-human animals have an ability to learn and process center embedding, a core property of human language syntax, is still debated. Artificial-grammar learning (AGL) has been used to compare humans and animals in the learning of center embedding. However, up until now, human participants have only included adults, and data on children, who are the key players of natural language acquisition, are lacking. We created a novel game-like experimental paradigm combining the go/no-go procedure often used in animal research with the stepwise learning methods found effective in human adults' center-embedding learning. Here we report that some children succeeded in learning a semantics-free artificial grammar with center embedding (A<sup>2</sup>B<sup>2</sup> grammar) in the auditory modality. Although their success rate was lower than adults', the successful children looked as efficient learners as adults. Where children struggled, their memory capacity seemed to have limited their AGL performance.

*Keywords:* artificial grammar; center embedding; children; go/no-go

## 1. Introduction

Language is one of the cognitive abilities that set humans and other animals apart. Therefore, theoretical inquiry into the nature of human language naturally has some biological significance (Jenkins 2000). Recent experimental studies show an even stronger flavor of biological make-up than previous theoretical linguistic studies, by directly comparing humans and animals or by studying animals' language learning ability (Abe and Watanabe 2011, Fitch and Hauser 2004,

*Editor:* Kleanthes K. Grohmann, University of Cyprus, Cyprus

*Received:* 23 August 2016

*Accepted:* 6 May 2020

*Published:* 19 August 2020

ISSN 1450–3417



CC BY 4.0 License  
© 2020 The authors

Gentner et al. 2006, Perruchet and Rey 2005, van Heijningen et al. 2009). These biolinguistic comparisons make use of artificial grammars, that is, very simple grammars artificially created for research purposes which can generate sentences by combining (often meaningless) words in a specific way. Artificial-grammar learning (AGL) has been used to study certain aspects of language learning in humans in a controlled way (Brooks et al. 1993, Knowlton and Squire 1996, Marcus et al. 1999, Morgan and Newport 1981, Reber 1967, Reber and Allen 1978). AGL paradigms have also been extended to study non-human animals' ability to learn some aspects of human language or other cognition (Gentner et al. 2006, Herman et al. 1984, Murphy et al. 2008, Rey et al. 2012, Savage-Rumbaugh et al. 1983).

The seminal work by Fitch and Hauser (2004) set the stage for a new trend of AGL. They asked whether the ability to learn a "context-free grammar" (or a phrase structure grammar in their terminology) was present in non-human primates (cotton-top tamarins). To properly characterize human-language sentences, we need context-free grammars (or grammars with more generative power such as context-sensitive grammars) (Chomsky 1957). In contrast, the description of natural behaviors of non-human animals seems to require only "finite-state grammars" (Fitch and Hauser 2004) or at best "the renewal process" (Kershenbaum et al. 2014), which have less generative power than context-free grammars. Both context-free grammars and finite-state grammars can generate surface strings such as "flying airplanes". In a native speaker's mind, however, such strings are represented as being hierarchically organized into units of phrases, as in [<sub>VP</sub> flying [airplanes]] or [<sub>NP</sub> [flying] airplanes]. Context-free grammars, being able to generate phrase markers, can capture this hierarchical nature of human language, while finite-state grammars cannot. Neither can the renewal process, which is characterized by a strong tendency to repeat elements and has properties somewhat between finite-state grammars and context-free grammars (Kershenbaum et al. 2014). Fitch and Hauser (2004) created two types of simple, meaningless artificial grammars, an  $A^nB^n$  grammar and an  $(AB)^n$  grammar, which they claimed had properties of a context-free grammar and a finite-state grammar, respectively.

Many AGL studies that followed used an  $A^nB^n$  grammar (or both of the above grammars) (Abe and Watanabe 2011, Bahlmann et al. 2006, Bahlmann et al. 2008, de Vries et al. 2008, Fedor et al. 2012, Friederici et al. 2006, Gentner et al. 2006, Hochmann et al. 2008, Lai and Poletiek 2011, Mueller et al. 2010, Perruchet and Rey 2005, Rey et al. 2012, Udden et al. 2012, van Heijningen et al. 2009). An  $A^nB^n$  grammar generates strings such as AB, AABB, and AAABBB, with A's and B's being represented by artificial words (in many cases, meaningless syllables). At the level of phrasal markers, these strings can be represented as [AB], [A [AB] B], and [A [A [AB] B] B], respectively. The latter two of these exemplify center embedding, which is sometimes said to be a hallmark of human language (for an opposing view, see Frank and Bod 2011). It is well known that center-embedding structure seen in a sentence such as "The boy [the girl liked] smiled" cannot be generated by finite-state grammars (Chomsky 1957; for a counterargument, see Christiansen and Chater 1999). So far, about 20 experiments using an  $A^nB^n$  grammar have been reported, which studied humans' and non-human animals'

learning of center embedding in an AGL paradigm (for a review, see Ojima and Okanoya 2014).

One critical problem of this line of research is that it is still unclear whether we can generalize the previous findings from human adult participants to human children, due to the lack of data. Human adults have high domain-general cognitive abilities as well as more domain-specific linguistic abilities. They may use both to carry out an AGL task, making it difficult to assess whether their success in AGL is attributable to the former, the latter, or both. Children's abilities are asymmetric in this respect; their domain-general cognitive abilities are not as high as those of adults, whereas they are more successful in language learning in the long run. To assess humans' abilities in AGL and compare them with other animals, we need data from both adults and children. Also, natural language acquisition takes place during childhood. It is children, not adults, who are the key players in humans' first language acquisition. This simple fact has been respected in other lines of AGL research, with many AGL experiments conducted on children as well as on adults (Braine 1963, Braine et al. 1990, Brooks et al. 1993, Saffran 2001, Saffran 2002), but the past  $A^nB^n$  studies have not tested human children. One recent brain-imaging study on infants (Winkler et al. 2018) used a mirror grammar which produced mirror sequences of pure tones such as A–B–C–B–A (1200Hz–1900Hz–1500Hz–1900Hz–1200Hz) and reported that infants' brain activity was sensitive to violations of regularities in such sequences. A mirror grammar is similar but not identical to an  $A^nB^n$  grammar (ABBA vs.  $A_xA_yB_yB_x$ ). For example, in mirror sequences of pure tones, the first A (ABCBA) is paired with the second A (ABCBAA) which is exactly the same pure tone as the first A, while in  $A^nB^n$  sequences of syllables, each A syllable is paired with a B syllable that is different from the A syllable itself. In the present study, we tried to provide data on children's learning of an  $A^nB^n$  grammar.

Here, we propose a unified experimental paradigm that could be used to test not only human adults and children but also other animals. The results of past  $A^nB^n$  studies have often been used to compare humans and other animals, but the experimental methodologies have been independently developed across human and animal studies and thus greatly differ. For example, the training of birds in AGL studies typically involves rewards and punishments (Gentner et al. 2006, van Heijningen et al. 2009), while humans' AGL does not (Bahlmann et al. 2008, Lai and Poletiek 2011). In the context of comparative cognitive studies, it is ideal that we use experimental paradigms that can be directly applied to many study subjects, including not only human adults and children but also non-human animals. Hence as a secondary aim of this study, we aimed to design an experimental paradigm that is applicable not only to human adults but also to children and other animals. We implemented the learning of an  $A^nB^n$  grammar in a go/no-go paradigm, which is often used to experimentally study non-human animals, with extensions towards future animal studies in mind.

Using this go/no-go paradigm, we ask two research questions. First, we ask whether the learning of a semantics-free artificial  $A^nB^n$  grammar is possible at all in human children. In other words, we ask whether there are any children who can succeed in this type of learning. In doing so, we target children in an age range such that they can be assumed to have a syntactic ability for center embedding in



their mother tongue, because the previous A<sup>n</sup>B<sup>n</sup> studies on human participants all targeted adults, who we can assume have a well-developed capacity for center embedding in their first languages. Native Japanese-speaking children aged 5 to 6 have a syntactic ability for center embedding, which is exemplified in grammatical constructions such as the embedding of complement clauses and adverbial clauses in Japanese (Kamio and Harada 1983). Hence, we targeted Japanese children who were 5 years of age or older, assuming that they had a basic syntactic ability for center embedding in their mother tongue. It is a different matter, however, whether they can master an artificial A<sup>n</sup>B<sup>n</sup> grammar without the aid of semantics. Past research (Fedor et al. 2012) has suggested that if there is no semantic information available, the learning of an A<sup>n</sup>B<sup>n</sup> grammar can be difficult even for adults. In the type of AGL paradigms we are adopting here, no semantic cues are available unlike in natural languages, which might make the learning and processing of center embedding even more difficult. Hence, we should first ask whether any child could ever succeed in our AGL task.

Expecting an affirmative answer to the first question, we secondly ask whether children's learning is more successful and efficient or less so than adults' (for definitions of success and efficiency, see Section 2.5). "Critical-period effects" or the long-term effects of age of immigration on the ultimate attainment of L2 (second language) morphosyntactic proficiency in immigrants are well known (DeKeyser 2000, Johnson and Newport 1989). These effects are such that naturalistic child L2 acquisition is more successful than adult L2 or foreign-language learning. Although these long-term effects are not directly relevant to our short-term AGL study, one electrophysiological comparison between infants and adults in the learning of an artificial "AXB" grammar (Mueller et al. 2012) also reports that adults were *less* successful than infants in learning non-adjacent dependencies. A follow-up study using the same paradigm has additionally reported that the learning of this AXB grammar is more difficult for 4-year-olds than for 2-year-olds (Mueller et al. 2019). The A<sup>n</sup>B<sup>n</sup> grammar used in the current study also generates non-adjacent dependencies. Hence one may expect that adults will be less successful in learning our A<sup>n</sup>B<sup>n</sup> grammar than children.

In contrast, short-term studies report adults' and older children's superiority to younger children. In the learning of English as a foreign language, the higher the Spanish-speaking learner's age was, the faster their morphosyntactic learning was (Muñoz 2006). In an AGL study, none of the youngest children (8-year-olds) succeeded in learning an artificial morphological rule, whereas adults and older children (12-year-olds) did (Ferman and Karni 2010). Moreover, adults were faster in learning than older children. In another AGL study, adults were more successful than children, achieving higher discrimination accuracy for predictive dependencies (Saffran 2001). These studies lead one to predict that adults will be more successful and more efficient (faster) in our AGL paradigm than children, which is completely the opposite of the earlier prediction. Our study was designed to provide further empirical evidence regarding the learning differences between adults and children.

In the experiment, we used a minimal A<sup>n</sup>B<sup>n</sup> grammar, perhaps the simplest one we could think of for our purposes, considering the previous observations that the learning of an A<sup>n</sup>B<sup>n</sup> grammar is very difficult for human adults (as

summarized in Ojima and Okanoya 2014) and the possibility that it is even more difficult for children. Hence our minimal  $A^nB^n$  grammar generated sentences only up to one level of embedding (AABB, but not AAABBB; Mueller et al. 2010) and used only four AB pairs.

To make children's learning of our  $A^nB^n$  grammar possible and easy, we not only followed but also extended the "starting small" procedure, which has been proven to be necessary for human participants to learn an  $A^nB^n$  grammar (Lai and Poletiek 2011). In this procedure, simple AB pairs without embedding, or 0-LoE (zero level of embedding) items, are learned first, that is, before AABB strings (1-LoE items) are presented. We extended this procedure and included cABc and AccB stages between the AB stage and the AABB stage. Also, the size of vocabulary was restricted to the bare minimum initially, to reduce the burden of lexical learning. Only at a later stage was the vocabulary size increased to generate a large set of novel sentences. By having these multiple graded stages, we could lead some children to the mastery of our  $A^nB^n$  grammar.

## 2. Materials and Methods

We tested human children aged 5 to 12, as well as adults, in the learning of a semantics-free  $A^nB^n$  grammar presented in the auditory modality. In particular, we asked whether the participants could extract the rules of center embedding from input sentences and apply them to new contexts. This study was approved by the ethics committee on experimental research involving humans of the Graduate School of Arts and Sciences at the University of Tokyo. Additional information on the methods can be found in Appendix A.

### 2.1. Participants

We analyze data from 38 participants. Half of them were young adults, and the other half children ( $n = 19$  in each group). The young adults (12 male) were students from four universities and were 20.5 years old on average (range 18.6–23.0,  $SD = 1.27$ ). The children (8 boys) were 9.0 years of age on average (range 5.4–12.5,  $SD = 1.76$ ), and most of them were primary school pupils. This wide age range was chosen to pinpoint the lowest age at which the learning of a semantics-free  $A^2B^2$  grammar is possible in our learning paradigm. All participants were healthy native speakers of Japanese and had never been diagnosed with any auditory, neurological, developmental, learning, or linguistic disorder. They were all right-handed, except for one child, who was left-handed (mean handedness quotient, 0.98 for adults, 0.91 for children; where 1 = completely right-handed and  $-1$  = completely left-handed; Oldfield 1971). We obtained written informed consent from all adult participants and the parents of all child participants. The children who had entered primary school provided written informed assent.

### 2.2. Stimuli

We created an  $A^2B^2$  grammar, with meaningless syllables in Japanese as words (words in the sense that the grammar put them together to make sentences, but

they were meaningless). “A” and “B” are the two main grammatical categories in our grammar. These categories are locational; A-category words always appear in the first half of the sentence, and B-category words always appear in the second half. “Words” in this grammar were all one syllable long. Four syllables (zo, re, ra, so) were designated as A words, and four other syllables (pi, bo, pa, nu) as B words. An additional syllable (kyu) was used as the sole c-category word. The mean duration of these syllables was 167.0ms (SD = 22.0ms). The A<sup>2</sup>B<sup>2</sup> grammar generated the following four sentence types: A<sub>x</sub>B<sub>x</sub>, cA<sub>x</sub>B<sub>y</sub>c, A<sub>x</sub>ccB<sub>x</sub>, and A<sub>x</sub>A<sub>y</sub>B<sub>y</sub>B<sub>x</sub>. AB sentences were employed because it has been known that human adults’ learning of an A<sup>n</sup>B<sup>n</sup> grammar is very difficult (or impossible in some cases) without AB strings having been learned beforehand (Lai and Poletiek 2011). In addition, we included cABc and AccB sentences, because their presence helped children analyze the internal structure of AABB sentences in the preliminary experiments.

Both grammatical and ungrammatical sentences were given as stimuli (Table 1). Ungrammatical sentences for the AB sentence type had the wrong pair of A and B (A<sub>x</sub>B<sub>y</sub>). This was also the case in ungrammatical sentences for the cABc and AccB sentence types (cA<sub>x</sub>B<sub>y</sub>c, A<sub>x</sub>ccB<sub>y</sub>). For the AABB sentence type, there were three types of ungrammaticality: single, swapped, and repetition. First, ungrammatical sentences of the single violation type had either the inner AB pair or the outer AB pair wrong (A<sub>x</sub>A<sub>y</sub>B<sub>y</sub>B<sub>z</sub> or A<sub>x</sub>A<sub>y</sub>B<sub>z</sub>B<sub>x</sub>). Second, those of the swapped violation type had the two original Bs swapped (A<sub>x</sub>A<sub>y</sub>B<sub>x</sub>B<sub>y</sub>). Lastly, those of the repetition violation type had a repetition of the same A followed by a repetition of the same B that was not paired with that A (A<sub>x</sub>A<sub>x</sub>B<sub>y</sub>B<sub>y</sub>). In the literature, there has been a concern that grammatical and ungrammatical A<sup>n</sup>B<sup>n</sup> strings can be discriminated by detecting repetitions (de Vries et al. 2008). Our stimuli cannot be discriminated accurately by this strategy alone, because both grammatical and ungrammatical stimuli contained repetitions.

We chose to present sentences in the auditory modality, because it was unlikely that all the child participants had fully acquired reading, to the level of adults. This choice made our study rare among the A<sup>n</sup>B<sup>n</sup> studies, a majority of which used visual presentation. Another study which used the auditory modality (Mueller et al. 2010) asked the participants to actively search for rules, whereas the present study did not. We used a speech synthesizer (AI Talk, AI Inc., Tokyo) to obtain audio files of the words spoken in a female voice. When these audio files were combined to obtain sentences, pauses of two different lengths were inserted. Short pauses of 40ms were inserted between the members of an adjacent pair, whereas long pauses of 480ms were inserted at boundaries of embedding. If we indicate a short pause by \_ and a long pause by \_\_, we can express the positions of these pauses for each sentence type as in the following: A\_B, c\_\_A\_B\_\_c, A\_\_c\_c\_\_B, and A\_\_A\_B\_\_B. Upon deciding these pause durations, we referred to Condition 4 of Mueller et al. (2010) and emphasized the central embedding in a similar way.

Stage	Stimulus type	Go stimuli	No-go stimuli
1	First AB	zo-pi, re-bo	zo-bo, re-pi
2	cABc	kyu-zo-pi-kyu, kyu-re-bo-kyu	kyu-zo-bo-kyu, kyu-re-pi-kyu
3	AccB	zo-kyu-kyu-pi, re-kyu-kyu-bo	zo-kyu-kyu-bo, re-kyu-kyu-pi
4	First AABB	zo-zo-pi-pi, zo-re-bo-pi, re-re-bo-bo, re-zo-pi-bo	zo-zo-pi-bo, zo-zo-bo-bo, zo-re-pi-bo, re-zo-pi-pi, re-zo-bo-bo, re-re-pi-bo
5	Probe AABB	(Same as above)	zo-zo-bo-pi, zo-re-pi-pi, zo-re-bo-bo, re-zo-bo-pi, re-re-pi-pi, re-re-bo-pi
6	Second AB	so-nu, ra-pa	so-pa, ra-nu
7	Second AABB	so-so-nu-nu, so-ra-pa-nu, ra-ra-pa-pa, ra-so-nu-pa	so-so-nu-pa, so-so-pa-nu, so-so-pa-pa, so-ra-nu-nu, so-ra-nu-pa, so-ra-pa-pa, ra-so-nu-nu, ra-so-pa-nu, ra-so-pa-pa, ra-ra-nu-nu, ra-ra-nu-pa, ra-ra-pa-nu
8	All AB	zo-pi, re-bo, so-nu, ra-pa	zo-bo, zo-nu, zo-pa, re-pi, re-nu, re-pa, so-pi, so-bo, so-pa, ra-pi, ra-bo, ra-nu
9	All AABB	zo-zo-pi-pi, zo-re-bo-pi, zo-so-nu-pi, zo-ra-pa-pi, & 12 others	zo-zo-bo-bo, zo-zo-nu-nu, zo-zo-pa-pa, zo-re-pi-bo, zo-re-bo-nu, zo-re-bo-pa, zo-re-nu-pi, zo-re-pa-pi, zo-so-pi-nu, zo-so-bo-pi, zo-so-nu-bo, zo-so-nu-pa, zo-so-pa-pi, zo-ra-pi-pa, zo-ra-bo-pi, zo-ra-nu-pi, zo-ra-pa-bo, zo-ra-pa-nu, & 54 others

Table 1: Stimuli in Language 1. Syllables coded in the same color constitute a valid pair; for example, the pair zo-pi, in red, is a grammatical sentence in Language 1, while the pair zo-bo, in red and blue, is not.

### 2.3. Stages

The AGL session consisted of multiple stages (Stage 1 to 9). Briefly, the complexity of the stimuli gradually increased from Stage 1 to 4 (AB → cABc → AccB → AABB), making it easier for participants to induce the center-embedding structure of AABB at Stage 4. Stage 5 checked the possibility that participants memorized ungrammatical sentences at Stage 4, without analyzing the internal structure of grammatical sentences (it is known in animal research that non-human animals sometimes focus on no-go stimuli only, to avoid punishments). Stage 6 presented new AB stimuli, which were different from Stage 1. At Stage 7, AABB stimuli created from these new AB stimuli were presented, to see if participants could apply the rule of center embedding learned at Stage 4 to new items straight away. Stage 8 looked at whether participants could handle four AB pairs (from Stages 1

and 6) simultaneously. Finally, at Stage 9 stimuli at full complexity (AABB created from all four AB pairs) were presented. At each stage, grammatical sentences (“go” stimuli) and ungrammatical sentences (“no-go” stimuli) were presented pseudo-randomly, each occupying 50% of the trials. No more than three sentences of the same grammaticality appeared in a row. Repetition of the same sentence in a row was restricted to twice.

One stage moved on to the next, when participants reached the accuracy criterion of 90% and underwent the minimum number of trials set specifically for that stage. The accuracy was calculated from the most recent 20 trials. The 90% accuracy criterion was adopted from Bahlmann et al. (2008) and Fedor et al. (2012), from which we also adopted some other methodological details. To ensure that we could compare children and adults, we chose the same accuracy criterion for children.

### 2.3.1. Stage 0: Practice

Participants were familiarized with the go/no-go procedure and the ways of responding at Stage 0, before engaging in AGL which started from Stage 1. At Stage 0, the Japanese words corresponding to “push” and “not push” were presented. After hearing “push”, they touched the red button. After hearing “not push”, they practiced refraining from touching the red button. After this initial familiarization, participants were given further practice trials with non-linguistic sounds, to get used to the system more. A pure tone was used as a go stimulus, and a white noise as a no-go stimulus.

### 2.3.2. Stages 1–3: First AB, cABc, AccB

At Stage 1, AB pairs were used as stimuli. There were only two go stimuli ( $A_xB_x$ : e.g., zo-pi) and only two no-go stimuli ( $A_xB_y$ : e.g., zo-bo; see Table 1). At Stage 2, the same AB pairs as Stage 1 were embedded between two c’s, as in cABc (e.g., kyu-zo-pi-kyu). At Stage 3, the same AB pairs as Stages 1 and 2 continued to be used but were separated by two c’s, as in AccB (e.g., zo-kyu-kyu-pi).

### 2.3.3. Stage 4: First AABB

The same two AB pairs as in the previous stages were used to create center-embedded AABB strings ( $A_xA_yB_yB_x$ : e.g., zo-re-bo-pi). The three types of ungrammatical AABB strings, namely, single ( $A_xA_yB_yB_y$ : e.g., zo-re-bo-bo), swapped ( $A_xA_yB_xB_y$ : e.g., zo-re-pi-bo), and repetition ( $A_xA_xB_yB_y$ : e.g., zo-zo-bo-bo), were presented at a ratio of 4:1:1. We had a larger number of single violations than swapped and repetition violations simply because no other combinations of A and B were possible for swapped and repetition types, given only two AB pairs.

### 2.3.4. Stage 5: Probe AABB

Novel *ungrammatical* AABB sentences were presented, with the four grammatical stimuli being the same as the previous stage. This was done to test the possibility

that at Stage 4 (First AABB), participants would focus only on no-go stimuli without learning the internal structure of go stimuli. If so, they would not be able to accurately classify novel no-go stimuli when all the familiar no-go stimuli were replaced with novel ones. Single, swapped, and repetition violations were presented at a ratio of 4:1:1, as at the previous stage. This stage finished regardless of the participant's accuracy when they had undergone 24 trials.

#### 2.3.5. Stage 6: Second AB

Two new AB pairs were presented. We had two grammatical AB strings (e.g., so-nu) and two ungrammatical AB strings (e.g., so-pa).

#### 2.3.6. Stage 7: Second AABB

AABB sentences (e.g., so-ra-pa-nu) were presented, which were created from the AB pairs introduced at the previous stage. Single, swapped, and repetition violations were presented at a ratio of 4:1:1.

#### 2.3.7. Stage 8: All AB

All AB pairs that had appeared previously were used. The go stimuli were the same as in Stages 1 and 6 (First AB and Second AB) (e.g., zo-pi), but many novel no-go stimuli (e.g., zo-nu) were used.

#### 2.3.8. Stage 9: All AABB

This stage presented AABB sentences created from all four AB pairs (e.g., zo-so-nu-pi). Single, swapped, and repetition violations were presented at a ratio of 2:1:1. With four AB pairs available, we could eliminate entirely, from the pool of ungrammatical stimuli, AABB strings which repeated the same A but used two different B's ( $A_xA_xB_xB_y$ : e.g., zo-zo-pi-bo) and those which repeated the same B after two different A's ( $A_xA_yB_yB_y$ : e.g., zo-re-bo-bo). Such ungrammatical sentences, which were a subset of single violations and were used at the other AABB stages, could be detected relatively easily by focusing on repetitions. By excluding this type of strings from the pool of ungrammatical stimuli, and by including repetition violations ( $A_xA_xB_yB_y$ : e.g., zo-zo-bo-bo), we ensured that the repetition counting strategy would not work at all (there were 96 grammatical and 96 ungrammatical sentences, and repetitions were contained in 24 sentences of each of these two types).

### 2.4. Procedure

Each participant was tested individually in a sound-proof room. A laptop computer with a touch-sensitive display controlled the presentation of auditory stimuli (sentences) and visual stimuli (scrambled pictures) and recorded the participant's behavioral responses, that is, touches on the display (Figure 1a). Touching was adopted as the mode of responding because non-human primates can easily do it and songbirds can produce a similar behavior, that is, pecking.

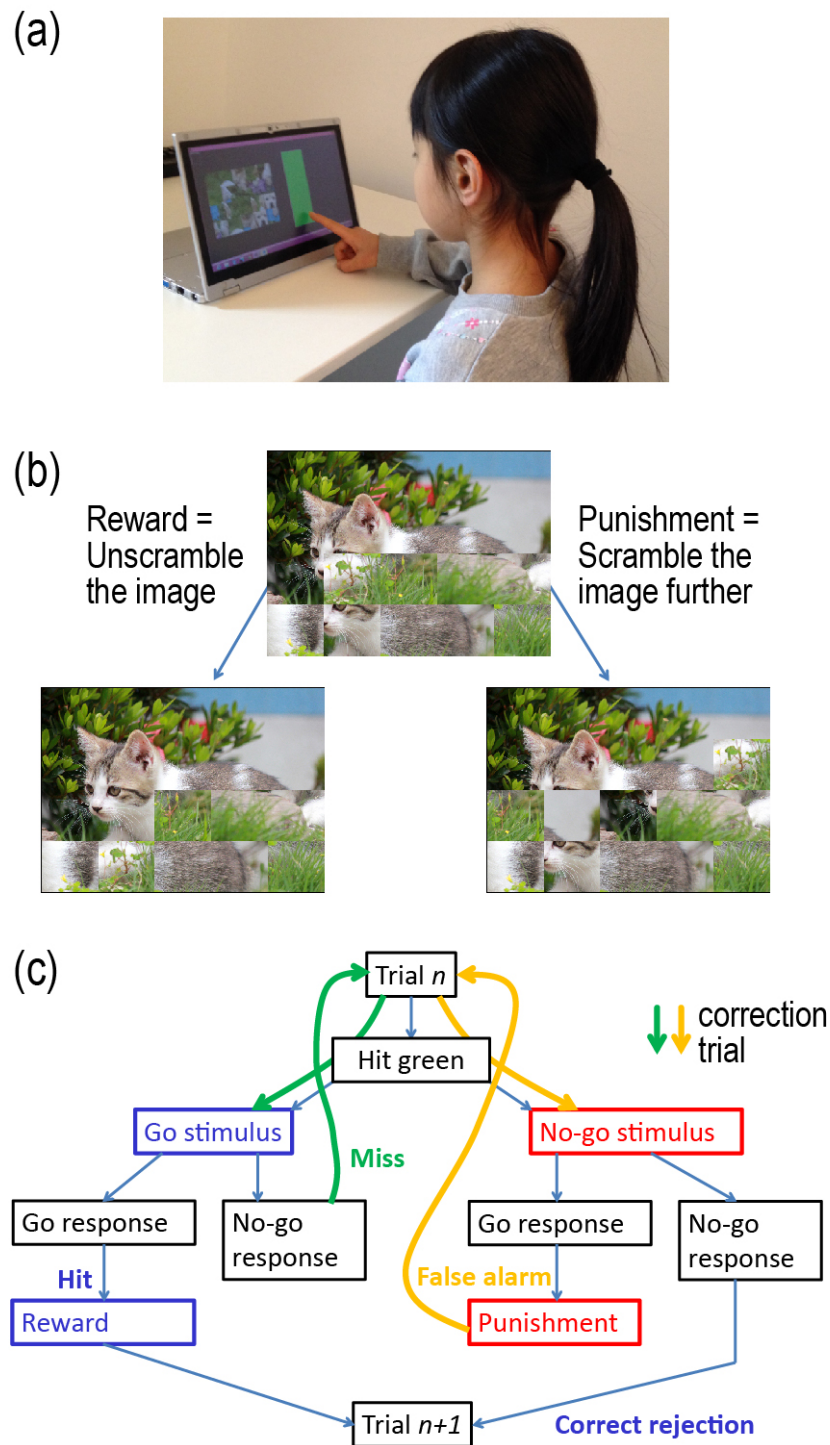


Figure 1: The image unscrambling task. (a) A child participant doing the task on a touch-sensitive display. (b) How a scrambled image got unscrambled as a reward or got scrambled further as a punishment. (c) A schematic description of the go/no-go procedure.

We employed a game-like task of unscrambling scrambled images. The computer displayed a scrambled picture on the left-hand side throughout the experiment. A green rectangle, which acted as a button, was shown on the right-

hand side. Touching the green button initiated a trial, which proceeded as follows. First, the green button disappeared. Eight hundred milliseconds later, an auditory stimulus (spoken sentence) was delivered. Upon the offset of the auditory stimulus, a red button appeared to the right of the original position of the green button. The red button remained on the display for a specific length of time, which was pre-determined for each stage (see below). While the red button was on, participants could touch it to indicate a "go" response. If they did not touch the red button, it was considered a "no-go" response. Grammatical stimuli required a "go" response, and ungrammatical ones required a "no-go" response. By touching or not touching the red button, participants tried to unscramble the scrambled picture (*Figure 1b*), which was the main task that they were overtly asked to do. After one scrambled image had gotten completely unscrambled, another scrambled image appeared. Our go/no-go procedure is schematized in *Figure 1c*.

We varied the duration of the red button depending on the length of the stimuli at that stage. The duration of the red button was 2s for Stages 1, 2, 6, and 8 which presented AB strings and cABc strings, 3s for Stage 3 which presented AccB strings, and 4s at Stages 4, 5, 7, and 9 which presented AABB strings. For more details about the go/no-go procedure see Appendix A.

There was a special procedure for Stage 1 (First AB). We needed to make certain that all participants were fully familiarized to the two grammatical AB pairs and became able to distinguish them from the ungrammatical ones. Loss of participants at this very first stage meant that we could not use their data at all to study the learning of center embedding, which was the primary aim of this study. However, preliminary experiments (whose data are not included here) had shown that the learning of two AB pairs from scratch was fairly difficult for some children; in some cases, no improvement was observed after an hour of training. To be able to test enough participants in the learning of center embedding in the main experiment, we made sure, in two ways, that all participants would learn the correct AB pairs at Stage 1. First, we simply let the participant try the task in the same way as the other stages. Second, as a last resort, if the participant (child or adult) could not discriminate between stimuli at all after 30 trials, we gave them a hint ("Some of these sounds are correct") or explicitly asked to focus on the two correct stimuli ("Why don't you remember this and this?"). One adult and nine children received this assistance. This way, we had all participants go to the next stage. The data of Stage 1 after 30 trials is hence contaminated in a sense and should not be taken at face value. Meaningful comparisons between adults and children begin at Stage 2.

## 2.5. Analysis

We analyzed participants' success rates and efficiency of AGL. In the experiment, participants proceeded to the next stage only if they had reached the accuracy criterion. The most obvious evidence for the mastery of our  $A^2B^2$  grammar is that they cleared all the stages. Hence, we statistically compared the adults and children in their success rates, or more precisely, in the numbers of individuals



who had succeeded in clearing all stages and those who had failed, by means of the chi-square test or Fisher's exact test (Sections 3.1 and 3.2).

Further, we assessed participants' efficiency of AGL, that is, how quickly they became able to discriminate between grammatical and ungrammatical stimuli, using two measures, which concern the first and the last part of each stage, respectively. Firstly, we analyzed the accuracy of the first 20 trials at each stage (Section 3.3). Because the accuracy was calculated from the most recent 20 trials, all participants had at least 20 data points at each stage but not necessarily any more. We separated those 20 trials into the first 10 trials and the second 10 trials and compared the two groups by means of the analysis of variance (ANOVA) with Group as a between-subject factor and Phase (first 10 vs. second 10 trials) as a within-subject factor. The dependent variable used as a measure of accuracy here was the percentage of accurate responses in the first or second 10 trials; a response was accurate if it was a go response to a go stimulus or a no-go response to a no-go stimulus. For each stage, this analysis included all participants who had undergone that stage whether they had succeeded or failed in clearing it. Secondly, we analyzed the number of trials that each participant had needed to finally reach the accuracy criterion of 90% (Section 3.4). In most cases, this variable was not normally distributed, which led us to use Mann-Whitney's U test to compare the two groups. This analysis excluded the participants who had not reached the accuracy criterion at a given stage (contra Fedor et al. 2012).

Where necessary, we did some additional statistical analyses that were appropriate for the data.

### 3. Results

#### 3.1. Overall Success Rates

A majority of the adult participants (16 out of 19) cleared all the stages including the last stage (All AABB stage) where many novel stimuli were presented, and thus demonstrated that they had learned our A<sup>2</sup>B<sup>2</sup> grammar (Figure 2). We also found five children who had succeeded in clearing all the stages. The youngest of these five children was 7.56 years of age. The success rate of the child participants was significantly lower than that of the adults (5/19 vs. 16/19,  $\chi^2(1) = 12.9$ ,  $p < .001$ ). Hence, the overall success rates suggest that as long as our A<sup>2</sup>B<sup>2</sup> grammar is concerned, adults' AGL is more successful than children's. We will look closely at each stage of our AGL task below.

#### 3.2. Stages

##### 3.2.1. Stage 0: Practice

All adult and child participants reached the 90% accuracy criterion and succeeded in clearing this stage easily, suggesting that our go/no-go paradigm and response inhibition required therein were not serious problems for the child participants.<sup>1</sup>

---

<sup>1</sup> Both groups performed accurately from the beginning (accuracy in the first 20 trials: 92.4% for adults and 93.5% for children,  $t < 1$ ).

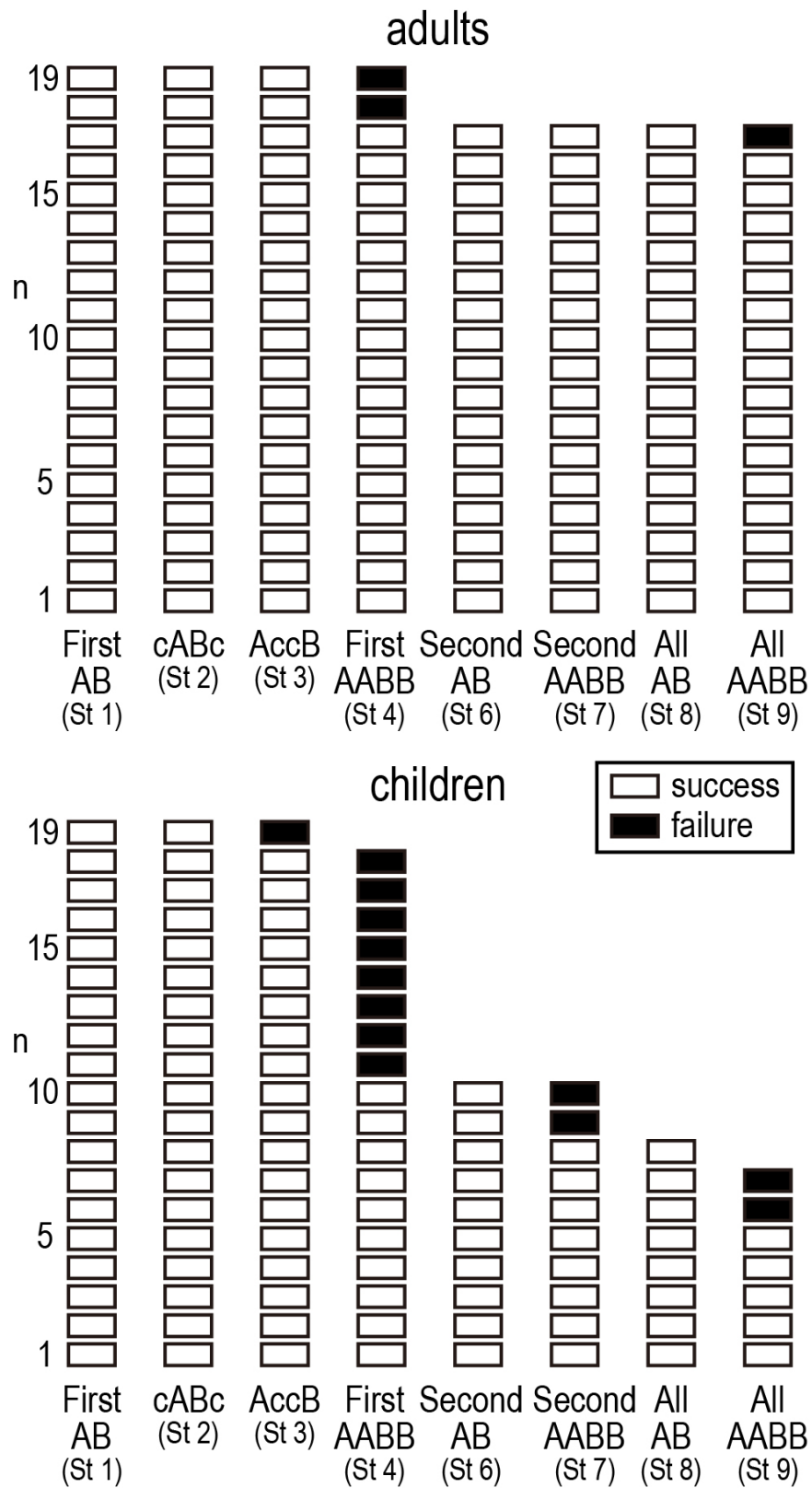


Figure 2: Number of participants who succeeded and those who failed at each stage (St = stage). Only the participants who succeeded in reaching the 90% accuracy criterion at a given stage proceeded to the next stage.

### 3.2.2. Stages 1–3: First AB, cABc, and AccB

After having successfully learned the first two pairs of AB at Stage 1, both adults and children easily extended this knowledge to the cABc sentences at Stage 2 (for accuracies of the first 20 trials at each stage, see *Figure 3* below). However, at the AccB stage (Stage 3), one child participant failed in reaching the accuracy criterion of 90%, being unable to apply the knowledge of the first AB to a non-adjacent context. She was the youngest participant (5.4 years old). Hence 18 of the original 19 children proceeded to Stage 4.

### 3.2.3. Stage 4: First AABB

A notable difference between adults and children was observed at this stage. Seventeen of the original 19 adult participants cleared this stage. One of the two who failed was the slowest adult at the First AB stage (Stage 1); it seems that this individual was poor at finding any kind of pattern across the board. Besides that, we did not notice anything special about these two participants. Among the 18 child participants, 10 succeeded and 8 failed. The success rate was significantly lower in children than in adults ( $p = .029$  by Fisher's exact test, two-tailed, which was used in place of the chi-square test due to the small number of adults who had failed). About half of the children (9 in total) left our AGL paradigm without showing evidence for the learning of our  $A^2B^2$  grammar.<sup>2</sup>

Here we further asked whether the participants had initially held the hypothesis that AABB strings were crossed ( $A_xA_yB_xB_y$ ), rather than center-embedded or nested ( $A_xA_yB_yB_x$ ). That would have resulted in the participants misinterpreting the swapped violations ( $A_xA_yB_xB_y$ ) as grammatical and the accuracies for them being lower than the other types of stimuli. This possibility was not supported by the data, in either the adults (accuracy for single violations  $66.7\% \pm 20.0$ , swapped  $70.5\% \pm 20.4$ , repetition  $75.9\% \pm 17.3$ , mean  $\pm$  SD, based on all trials) or in the children (single  $58.6\% \pm 14.4$ , swapped  $62.6\% \pm 19.2$ , repetition  $74.0\% \pm 20.0$ ). This tendency remains similar even if we restrict the analysis to the children who failed in clearing this stage (single  $49.2\% \pm 10.6$ , swapped  $63.0\% \pm 21.4$ , repetition  $70.7\% \pm 18.9$ ). Hence, we did not obtain clear evidence that the participants had had the "crossed" hypothesis.

However, it was not the case that the participants had treated all types of stimuli equally. A repeated-measures ANOVA with Grammaticality as a within-subject factor and Group as a between-subject factor revealed that grammatical stimuli were judged more accurately (83.4%) than ungrammatical ones (66.7%), regardless of Group (Grammaticality,  $F(1, 35) = 24.9$ ,  $p < .001$ ; Grammaticality  $\times$  Group,  $F < 1$ ; Group,  $F(1, 35) = 2.58$ ,  $p = .117$ ; accuracies calculated from all trials). In the participants who experienced five or more examples for all of the three

<sup>2</sup> Among the 18 children who underwent the First AABB stage, eight had been given assistance at the First AB stage (Stage 1). Four of them cleared the First AABB stage, whereas the other four failed. This ratio was similar to the children who had not been given help at the First AB stage; six of them cleared the First AABB stage, while the remaining four failed. Hence about 50% of the children could not clear the First AABB stage whether or not they had received help at the First AB stage.

types of violations (single, swapped, repetition), a repeated-measures ANOVA showed a significant main effect of Ungrammaticality Type (three levels,  $F(2, 28) = 9.62$ ,  $p = .001$ ). Bonferroni-corrected multiple comparisons revealed that repetition violations had been judged significantly more accurately than single violations ( $p = .002$ ).

Single violations, or the most difficult violations as shown above, had a mismatch in either the inner or outer AB pair. To see whether this factor (inner vs. outer) had any effects, a repeated-measures ANOVA was run on the accuracies of single violations, with Side (inner vs. outer) as a within-subject variable and Group (children vs. adults) as a between-subject variable. Neither the main effects nor the interaction reached significance (all  $p$ 's  $> .1$ ; accuracies, inner  $51.1\% \pm 21.1$  vs. outer  $59.2\% \pm 19.7$  in adults; inner  $52.2\% \pm 24.6$  vs. outer  $62.3\% \pm 19.0$  in children; data from participants who had 5 or more trials for both types), although mismatches in inner pairs were numerically less accurate.

#### 3.2.4. Stage 5: Probe AABB

Upon encountering novel no-go stimuli, the adult participants continued accurate discrimination (mean accuracy  $\pm$  SD,  $93.6\% \pm 12.0$ ), suggesting that they had not relied on the rote memorization of no-go stimuli at the First AABB stage. Similarly, the children who had reached the accuracy criterion of 90% at Stage 4 (First AABB) could discriminate the familiar go stimuli and the novel no-go stimuli accurately ( $83.3\% \pm 12.3$ ) and did not differ significantly from the adults ( $t(17) = 1.63$ ,  $p = .122$ ).<sup>3</sup> Since this stage contained only 24 trials and ended regardless of the accuracy, it is not included in the analyses presented below (accuracy in the first 20 trials in Section 3.3 and number of trials needed to clear a stage in Section 3.4).

#### 3.2.5. Stages 6 and 7: Second AB and Second AABB

In both groups of participants, all individuals (17 adults and 10 children) succeeded in learning the second AB pairs, at the Second AB stage. If participants had learned the rule of embedding at the First AABB stage, then they would be able to apply this rule to novel AABB strings made from new AB pairs, at the Second AABB stage. All adult participants responded in accordance with this prediction. Eight of the 10 child participants did so, too ( $p = .128$  by Fisher's exact test, two-tailed). Both adults and children performed quite accurately from the beginning (adults,  $90.0\% \pm 11.9$  for mean accuracy  $\pm$  SD in the first 10 trials; children,  $87\% \pm 10.0$ ).

#### 3.2.6. Stage 8: All AB

The four grammatical sentences (four AB pairs) at this stage had been used at previous stages, Stage 1 (First AB) and Stage 6 (Second AB). The adult participants had no problem discriminating the grammatical from the ungrammatical

<sup>3</sup> Data at this stage could be obtained for only a subset of the participants (13 adults and 6 children) due a technical problem.

sentences including novel ones ( $81.8\% \pm 15.0$  in the first 10 trials). All the eight children who had remained also succeeded, but many seem to have struggled ( $66.3\% \pm 16.5$  in the first 10 trials), as will be clear later in analyses of the numbers of trials needed to reach the accuracy criterion.

### 3.2.7. Stage 9: All AABB

All adult participants but one succeeded in reaching the accuracy criterion and thus showed that they were able to discriminate between grammatical and ungrammatical sentences. Half of the grammatical and two thirds of the ungrammatical sentences were novel sentences that had not appeared at the previous stages. Despite this, the adult participants were nearly 90% accurate from the beginning ( $88.3\% \pm 11.5$  in the first 10 trials). It is unlikely that their performance at this stage was based purely on rote memory; even if they had memorized all the strings from previous stages, they would have got only 41.7% of the stimuli correct.

Seven child participants tried this stage (there was one child who had cleared the previous stage but could not participate in the final stage because he used up all the time permitted by the ethics committee to clear the previous stages). Five succeeded in reaching the accuracy criterion. Given that the successful children correctly judged a great number of novel sentences at this stage, we can exclude the possibility that they had memorized all those sentences and discriminated sentences based on rote memory. It is also true that most child participants including those who succeeded struggled with discrimination at this stage ( $74.3\%$  accurate  $\pm 10.5$  in the first 10 trials,  $64.3\% \pm 13.0$  in the second 10 trials). This is likely to be because it had already been difficult for them to discriminate between grammatical and ungrammatical AB's at the previous stage (Stage 8, All AB). The difference in age between those children who had cleared all the stages and those who had not was not statistically significant (mean  $\pm$  SD, 9.63 years old  $\pm 2.16$  vs.  $8.73 \pm 1.52$ ,  $t(17) = .956$ ,  $p = .352$ ).

To better understand how the successful children discriminated AABB stimuli at the All AABB stage, we considered the possibility that they had performed differently on the three types of ungrammatical stimuli (single, swapped, repetition). A repeated-measures ANOVA revealed no significant difference among the three violation types ( $F(2, 8) < 1$ ; average accuracy, 73.6% for single violations, 73.1% for swapped violations, and 79.5% for repetition violations). Here it is particularly important to note that the children's performance on swapped violations (A1–A2–B1–B2) was not worse than the other two types of violations. The successful children not only learned that some syllables must appear together (e.g., A1 and B1) but also found out that correct AABB strings had a center-embedding structure and thus were able to reject swapped violations which had the correct elements but not a center-embedding structure.

In a post-experiment verbal report, in which participants were asked to describe in their words how they had discriminated stimuli, all individuals (adult or child) who succeeded except one adult said they had checked the outer and

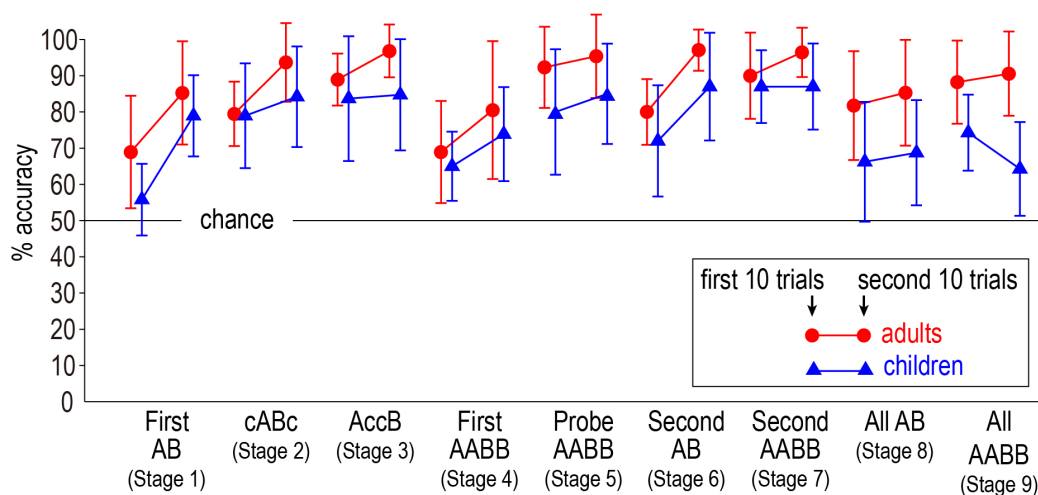
inner pairs for AABB strings. The one exceptional adult said that she had relied on intuition.

### 3.3. Discrimination Accuracy for the First 20 Trials

The accuracies of discrimination in the first 20 trials are shown in *Figure 3*. It is clear from the figure that both adults and children showed an above-chance performance even in the first 10 trials at most stages. This suggests that the participants could efficiently apply the previously acquired knowledge to a new stage, for example, apply the knowledge of the first two AB pairs acquired at the First AB stage to the cABc stage. Also, within the first 20 trials, their performance improved quickly at some stages.

ANOVAs showed that the main effect of Phase (first 10 vs. second 10 trials) was statistically significant at the following stages (see *Table 2*): First AB, cABc, First AABB, and Second AB. Accuracies in the second 10 trials were better than those in the first 10 trials at these stages unanimously. *Figure 3* additionally shows a clear tendency that the adults' performance was already better in the first 20 trials than the children's. ANOVAs revealed a significant main effect of Group (higher accuracies for adults) at the First AB, AccB, Second AB, All AB, and All AABB stages (*Table 2*). The interaction between Group and Phase was significant at the cABc and All AABB stages. At the cABc stage, the two groups did not differ significantly in the first 10 trials ( $t(36) < 1$ ), but in the second 10 trials, the adults were more accurate ( $t = 2.28, p = .029$ ). At the All AABB stage, the adults were more accurate both in the first and second 10 trials than were children ( $t(22) = 2.65$  and  $4.67, p = .015$  and  $< .001$ , respectively).

It should be stressed that the stages where there was no significant group difference were the critical First AABB and Second AABB stages. At the First AABB stage, the participants encountered AABB strings for the first time, and both groups seem to have had difficulty in discrimination initially. At the Second AABB stage, both groups were about 90% accurate from the beginning, suggest-



*Figure 3: Accuracies of the first 20 trials at each stage. In those 20 trials, the first 10 trials and the second 10 trials are shown separately. The values at a given stage are based on both the participants who succeeded and those who failed in reaching the accuracy criterion at that stage. Error bars, SD.*

Stage	Stimuli	d.f.	Group		Phase		Group × Phase	
			F	p	F	p	F	p
1	First AB	1, 36	9.58	.004	44.0	< .001	1.35	
2	cABc	1, 36	1.88		28.6	< .001	6.04	.019
3	AccB	1, 36	5.89	.020	4.09	.051	2.39	
4	First AABB	1, 35	1.84		12.3	.001	< 1	
6	Second AB	1, 25	5.32	.030	47.5	< .001	< 1	
7	Second AABB	1, 25	2.84		2.55		2.55	
8	All AB	1, 23	9.09	.006	< 1		< 1	
9	All AABB	1, 22	19.1	< .001	1.72		4.49	.046

Table 2: Results of ANOVAs for the accuracies in the first 20 trials. The factor Phase had two levels: First and second 10 trials at each stage. Only *p*-values smaller than .1 are shown.

ing that not only the adults but also the children could efficiently apply the rule of center embedding that they acquired from the First AABB stage to the two new pairs of AB.

### 3.4. Number of Trials Needed

Analyses of the number of trials needed to reach the accuracy criterion favored the adults in some but not all cases. As shown in *Figure 4*, there was a tendency for the adults to have needed significantly fewer trials to reach 90% accuracy, compared with the children, at the AccB, Second AB, All AB, and All AABB stages ( $p = .003, .009, .031, .000$ , respectively, by Mann-Whitney's U test). Note that the adults and the children did not differ statistically at the First AABB and the Second AABB stage; at these stages, the non-significant numeric difference goes in the opposite direction, with the children having needed fewer trials.

It is also noteworthy that the children needed a large number of trials to clear the All AB stage (more than 50 trials on average; *Figure 4*), despite the fact that they had previously learned all the grammatical sentences and those sentences were simple AB strings without embedding. Perhaps this suggests that it was difficult for them to handle four arbitrary pairs simultaneously. Given this difficulty, it may be natural for the children to have needed many trials to clear the All AABB stage, where the adults needed just about 20 trials.

### 3.5. Additional Experiment

We ran an additional experiment and confirmed that it was extremely difficult (and perhaps impossible) to clear the last stage (All AABB) if the four AB pairs had not been learned in advance. After the practice stage, the participants directly entered AABB stages (First AABB, Second AABB, All AABB, in this order), without learning the AB pairs. None of the five participants could reach 90% accuracy during 100 trials of the All AABB stage, although four participants

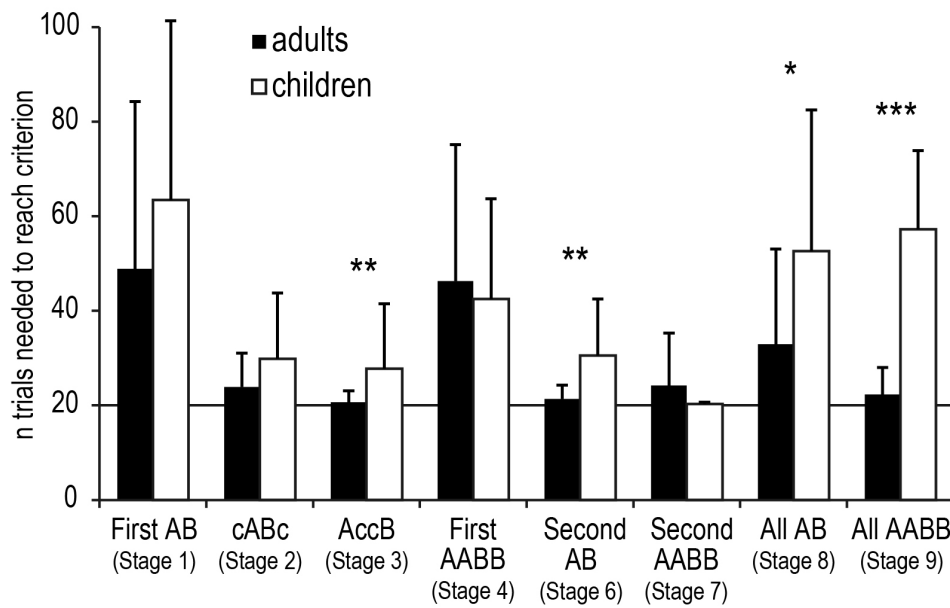


Figure 4: Numbers of trials needed to reach the 90% accuracy criterion at each stage. The minimal number possible is 20 here. Asterisks show statistically significant differences between the groups (\*  $p < .05$ , \*\*  $< .01$ , \*\*\*  $< .001$ ). Error bars, SD.

cleared either the First AABB stage or the Second AABB stage and the remaining one cleared both these stages.<sup>4</sup>

#### 4. Discussion

In a game-like go/no-go paradigm, young adults and children discriminated spoken strings which were consistent or not consistent with a minimal  $A^nB^n$  grammar. Most of the adult participants cleared all the stages of the experiment, showing clear signs of the mastery of the grammatical rules. It is only less than a third of the children who cleared all stages, which suggests that children were less likely to succeed in ultimately mastering the grammar. However, the successful children were as fast as adults at the critical First AABB (Stage 4) and Second AABB stage (Stage 7), where they were required to induce the rule of center-embedding for the first time (First AABB) and apply this rule to a different pair of AB (Second AABB). The children were slower at the All AABB stage (Stage 9) than the adults, but it seems to be due to the difficulty they had in handling four

<sup>4</sup> Five right-handed, healthy university students, who had not taken part in the main experiment, took part after providing written consent. They proceeded to the next stage if they had undergone 100 trials or reached the 90% accuracy criterion. In other aspects of the procedure, this additional experiment was identical to the main experiment. The result accords well with the previously reported observation that the induction of an  $A^nB^n$  grammar is extremely difficult if input is not given in a staged manner, that is, if the participant is not given AB pairs before AABB strings (Lai and Poletiek 2011). In a post-experiment verbal report, none of these five participants said that they had relied on a rule to discriminate stimuli. Instead they tried to memorize the go stimuli (without knowing the internal structure), which worked at the First AABB stage and the Second AABB stage where there were only four go stimuli, but did not work at the All AABB stage where there were 16 go stimuli (and 72 no-go stimuli).



AB pairs simultaneously. This was already evident at the All AB stage (Stage 8). Hence the children's poor performance at the All AABB stage does not necessarily indicate their weakness in rule application. Rather, it may have resulted from limitations on their memory capacity. We discuss these and other points in more detail below.

The results show that at least some of the children who took part succeeded in learning our  $A^2B^2$  grammar. The youngest child who succeeded was 7 years of age. Hence, children who are 7 or older possess a linguistic ability that enables the learning and processing of center embedding in a semantics-free artificial grammar, at least up to one level of embedding. Language acquisition research has generally suggested that semantics plays facilitative roles. Children's first language acquisition seems to benefit from semantic and pragmatic information (Pinker 1984, Pinker 1987, Tomasello and Akhtar 1995). AGL is also easier when there is more semantic information available to the learner (Moeser and Bregman 1973, Mori and Moeser 1983). More specifically, semantics greatly helps adults' learning of center embedding in an artificial grammar (Fedor et al. 2012). While acknowledging these beneficial roles of semantics in natural and artificial language learning, we suggest that children do possess a linguistic ability to learn center embedding without the aid of semantics, when the level of embedding is one. It will be ideal if this line of research can be extended to two levels of embedding (i.e., an  $A^3B^3$  grammar), to ensure the generalizability of the results. However, we would also like to point out that an  $A^3B^3$  grammar, with two levels of embedding, has been tested only in the *visual* modality so far.

Our data from children constitute important evidence that the ability to learn an artificial  $A^nB^n$  grammar originally observed in adults also exists in children to some extent. The learning of center embedding in an artificial grammar does not depend on the high domain-general cognitive abilities that are available only to adults; it depends on some abilities that are shared by adults and children. With this evidence available, we are now in a better position than before, to make meaningful comparisons between humans and non-human animals in the learning of an  $A^nB^n$  grammar.

A significantly smaller portion of the child participants succeeded in clearing all the stages, compared to the adult participants. This is largely due to the failure of many children at the First AABB stage, where they had to discover the rules of the grammar for the first time. Children's lower success rate is consistent with the past studies which reported adults and older children's initial (but not necessarily long-term) advantages over younger children in AGL (Ferman and Karni 2010, Saffran 2001). It remains unclear why children are limited in their capacity to induce the rules of our  $A^2B^2$  grammar, but one possibility is that they are not as good at detecting grammatical patterns *consciously* as adults. Almost all successful participants verbally reported that they had checked the inner and outer pairs of AB in AABB strings, suggesting that they consciously understood the rule of center embedding in our grammar. In contrast, infants' remarkable ability to learn statistical patterns, algebraic rules, an AXB grammar and a mirror grammar, does not seem to require consciousness (Marcus et al. 1999, Mueller et al. 2012, Saffran et al. 1996, Winkler et al. 2018). More

implicit, less conscious ways of learning that have rarely been applied to an  $A^nB^n$  grammar (Udden et al. 2012) may be worth testing in the future.

Although a lower rate of children succeeded in grammar induction at the First AABB stage, the group difference was not significant at this stage in how fast their performance improved (that is, in the accuracies of the first 20 trials and in the number of trials needed to reach the accuracy criterion). Also, at the Second AABB stage, children could apply the grammar they had induced at the First AABB stage to two new pairs of AB, as efficiently as adults. The children's discrimination performance was already around 90% accurate in the first 10 trials of the Second AABB stage (Figure 3), and the successful children needed only about 20 trials to reach the accuracy criterion at this stage (Figure 4). Thus, it seems that once grammar induction had gone successfully, grammar application also went efficiently, as long as the number of AB pairs involved was limited to two.

On the other hand, children had greater difficulty in handling four AB pairs simultaneously, than did adults. Even though they had already been familiarized with all four AB pairs before, their initial performance at the All AB stage was fairly low, that is below 70% accuracy on average (Figure 3), and significantly lower than the adults' performance. This is also true for the All AABB stage; they had problems applying the grammar to four AB pairs quickly. Also, the children who succeeded in reaching the criterion at the All AABB stage needed a far larger number of trials to do so than did the adults, who required slightly more than the minimum of 20 trials (Figure 4). Children's difficulty in handling four AB pairs may be due to limitations on their memory capacities. Then their difficulty at the All AABB stage cannot be attributed to their grammar application ability *per se*.

We developed a novel experimental paradigm which combined the go/no-go procedure used in animal research and the stepwise presentation of input used in human research. We believe that we can apply our paradigm to non-human animals, if we replace the reinforcement procedure (image unscrambling) with the kind of reinforcement compatible with animals in question, for example, giving food as a reward. However, our adoption of a go/no-go procedure may have introduced one complication. As in typical go/no-go procedures, when participants made a go response to a no-go stimulus, they got punished (the image got scrambled further) in our experiment. Essentially this is so-called negative evidence in language learning. In natural conversations with their children, parents do not provide explicit feedback on the ungrammaticality of the children's utterances (Brown and Hanlon 1970). Hence children's first language acquisition succeeds given positive evidence (grammatical sentences in that language), without negative evidence (information about which sentences are ungrammatical). Given that natural language acquisition does not rely on negative evidence, our go/no-go procedure may be unnatural in this respect.

Considering the shortcoming noted above, one direction for future comparisons across ages (e.g., adults vs. children) and across species (e.g., humans vs. birds) is to make a better use of the study subject's spontaneous behavior, and not to use reinforcement, punishment, and/or negative evidence. Two studies reporting non-human animals' sensitivity to  $A^nB^n$  patterns took advantage of the animals' spontaneous behavior (Abe and Watanabe 2011, Rey et al. 2012). These and other studies, including the one that used the implicit mode

of learning (Udden et al. 2012) discussed earlier, might offer methodological suggestions for how to design an experiment that is applicable to a wide range of study subjects.

## 5. Conclusion

We used a minimal  $A^2B^2$  grammar to see if human children aged 5 to 12, as well as adults, could learn a semantics-free artificial grammar with one level of center embedding, in the auditory domain. This AGL task was implemented in a go/no-go paradigm often used in animal research. Results showed that some children succeeded in the learning of this grammar. They not only extracted the rules of center embedding from input sentences but also applied them to new contexts. Our data are consistent with the view that human children have a latent ability to learn and process center embedding without the aid of semantics, at least up to one level of embedding. Hence human adults' ability to learn an  $A^nB^n$  grammar reported by previous research does not depend on the high domain-general cognitive abilities possessed only by adults. The clearest difference between adults and children emerged at the stage where they were first required to discover the rule of center embedding (First AABB stage). Fewer children succeeded in this than adults, but the successful children were as efficient as the adults in the induction (First AABB stage) and application of the grammatical rule to new elements (Second AABB stage). Children struggled at the final stage (All AABB stage), but that seemed to be due to memory constraints, not rule application capacity *per se*. These data provide a new insight into children's AGL ability and motivate further methodological innovations to study adults, children, and animals in an identical manner.

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP23240033, JP23520757, JP16K02959, JP19H01280, Grant-in-Aid for Scientific Research on Innovative Areas #4903 (Evolinguistics) Grant Number JP18H05065, and ERATO, Japan Science and Technology Agency. We thank all the participants and the parents who accompanied the child participants. We also thank Andrew McNulty and two anonymous reviewers for valuable comments.

## Appendix A

### *Two Languages*

We prepared two “languages”, Language 1 and Language 2, which shared the basic grammatical rules but used different AB pairs, to exclude the possibility that learning would depend on particular combinations of A and B words. For example, A word “zo” is paired with B word “pi” in Language 1, whereas in Language 2, it is paired with “bo”. Each participant was tested in either one of these languages. The correct pairs were as follows: zo–pi, re–bo, so–nu, and ra–pa in Language 1, and zo–bo, re–pi, so–pa, and ra–nu in Language 2.

### *Scrambled Pictures*

Visual stimuli were used as part of the reinforcement procedure in our go/no-go paradigm. These stimuli were pictures of things young children are generally fond of, including animals, fruits, sweets, and trains. Each picture was presented initially in a “scrambled” form (*Figure 1b*). As the participant's learning proceeded, scrambled pictures got gradually unscrambled as a reward.

### *The Go/No-go Procedure*

Touching the red button related to unscrambling in the following way (*Figure 1c*). Touching (a go response) for a “go” stimulus, that is, a “hit”, led to unscrambling by one block; that is, one block returned to its original position, as a reward for a successful behavior. Touching for a “no-go” stimulus, that is, a “false alarm”, led to further scrambling by two blocks (rather than unscrambling), as a punishment for an unsuccessful behavior. Feedback (unscrambling or further scrambling) was given immediately after the participant's response. After a false alarm, the same auditory stimulus was presented again, for a correction trial. If the participant did not touch the red button (a no-go response), the image remained the same, whether the stimulus was a go stimulus or a no-go stimulus. However, not touching for a go stimulus, that is, a “miss”, led to the repetition of the same auditory stimulus at the next trial (correction trial), whereas not touching for a no-go stimulus, that is, a “correct rejection”, led to a different auditory stimulus at the next trial. Correction trials were excluded from calculations of accuracies; that is, accuracies were based only on novel trials.

Participants were *not* told beforehand about these rules of how their responses related to the functioning of the system. To unscramble images, they had to discover these rules by themselves, and had to get as many hits as possible while avoiding false alarms. Misses led to correction trials and thus caused delays, whereas correct rejections did not. To proceed smoothly, participants had to avoid misses.

### *Limit of 100 Trials*

In addition to the accuracy criterion of 90%, at Stages 4 (First AABB), 7 (Second AABB), and 9 (All AABB), where participants were trained and tested on AABB

strings, an upper limit of 100 was imposed on the number of trials (although training continued beyond this as long as the accuracy remained 75% or higher, to see whether the performance would keep improving to soon reach 90% or rather it would go down to below 75% without showing clear evidence of mastery). Hence participants finished these stages (4, 7, and 9) either when they had reached the accuracy criterion (90%) or when they had undergone 100 trials without reaching the criterion.

### *Review Trials*

In addition to normal trials, “review” trials were also given, on which participants were tested on items that they had already learned at the previous stages. Review trials were used at Stages 2, 3, 4, and 7 (cABc, AccB, First AABB, Second AABB) and were interleaved with normal trials. The ratio of normal vs. review trials was 4:1 at these stages. At Stage 2 (cABc), review trials presented the AB pairs they had already experienced at Stage 1 (First AB). The same review items (AB stimuli) were also used at Stage 3 (AccB). The review items at Stage 4 (First AABB) were cABc and AccB strings from Stages 2 and 3. At Stage 7 (Second AABB), review trials presented the AB pairs used for Stage 6 (Second AB). Review trials did not enter into calculations of accuracy or the number of trials needed to reach the accuracy criterion.

### *Other Details*

The minimum number of trials that a participant had to undergo was as follows: 15 for Stage 0, 30 for Stage 1, 20 for Stages 2, 3, and 4, 24 for Stage 5, 40 for Stage 6, 20 for Stages 7, 8, and 9.

The order of trials was specified in a list. We prepared one list for each stage and used the same list for all participants, who nonetheless started from different positions; that is, if a participant finished at trial  $i$ , the next participant began from trial  $i+1$ .

## References

- Abe, Kentaro & Dai Watanabe. 2011. Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience* 14, 1067–1074. doi:10.1038/nn.2869
- Bahlmann, Jörg, Thomas C. Gunter & Angela D. Friederici. 2006. Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience* 18, 1829–1842. doi:10.1162/jocn.2006.18.11.1829
- Bahlmann, Jörg, Ricarda I. Schubotz & Angela D. Friederici. 2008. Hierarchical artificial grammar processing engages Broca's area. *NeuroImage* 42, 525–534. doi:10.1016/j.neuroimage.2008.04.249
- Braine, Martin D. S. 1963. On learning the grammatical order of words. *Psychological Review* 70, 323–348. doi:10.1037/h0047696
- Braine, Martin D. S., Ruth E. Brody, Patricia J. Brooks, Vicki Sudhalter, Julia A. Ross, Lisa Catalano & Shalom M. Fisch. 1990. Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language* 29, 591–610. doi:10.1016/0749-596X(90)90054-4
- Brooks, Patricia J., Martin D. S. Braine, Lisa Catalano, Ruth E. Brody & Vicki Sudhalter. 1993. Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language* 32, 76–95. doi:10.1006/jmla.1993.1005
- Brown, Roger & Camille Hanlon. 1970. Derivational complexity and order of acquisition on child speech. In John R. Hayes (ed.), *Cognition and the Development of Language*, 11–53. New York, NY: Wiley.
- Chomsky, Noam. 1957. *Syntactic Structures*. The Hague, The Netherlands: Mouton.
- Christiansen, Morten H. & Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23, 157–205.
- de Vries, Meinou H., Padraic Monaghan, Stefan Knecht & Pienie Zwitserlood. 2008. Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition* 107, 763–774. doi:10.1016/j.cognition.2007.09.002
- DeKeyser, Robert M. 2000. The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition* 22, 499–533.
- Fedor, Anna, Máté Varga & Eörs Szathmáry. 2012. Semantics boosts syntax in artificial grammar learning tasks with recursion. *Journal of Experimental Psychology: Learning Memory and Cognition* 38, 776–782. doi:10.1037/a0026986
- Ferman, Sara & Avi Karni. 2010. No childhood advantage in the acquisition of skill in using an artificial language rule. *PLoS ONE* 5:e13648. doi:10.1371/journal.pone.0013648
- Fitch, W. Tecumseh & Marc D. Hauser. 2004. Computational constraints on syntactic processing in a nonhuman primate. *Science* 303, 377–380. doi:10.1126/science.1089401

- Frank, Stefan L. & Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22, 829–834. doi:10.1177/0956797611409589
- Friederici, Angela D., Jörg Bahlmann, Stefan Heim, Ricarda I. Schubotz & Alfred Anwander. 2006. The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America* 103, 2458–2463. doi:10.1073/pnas.0509389103
- Gentner, Timothy Q., Kimberly M. Fenn, Daniel Margoliash & Howard C. Nusbaum. 2006. Recursive syntactic pattern learning by songbirds. *Nature* 440, 1204–1207. doi:10.1038/nature04675
- Herman, Louis M., Douglas G. Richards & James P. Wolz. 1984. Comprehension of sentences by bottlenosed dolphins. *Cognition* 16, 129–219.
- Hochmann, Jean-Rémy, Mahan Azadpour & Jacques Mehler. 2008. Do humans really learn  $A^nB^n$  artificial grammars from exemplars? *Cognitive Science* 32, 1021–1036. doi:10.1080/03640210801897849
- Jenkins, Lyle. 2000. *Biolinguistics: Exploring the Biology of Language*. Cambridge, United Kingdom: Cambridge University Press.
- Johnson, Jacqueline S. & Elissa L. Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21, 60–99.
- Kamio, Akio & Kazuko I. Harada. 1983. A repetition experiment on children's comprehension of complex sentences in Japanese. In Shiro Hattori & Kazuko Inoue (eds.), *Proceedings of the 13<sup>th</sup> International Congress of Linguists*, 772–775. Tokyo, Japan: Proceedings Publishing Committee.
- Kershenbaum, Arik, Ann E. Bowles, Todd M. Freeberg, Dezhe Z. Jin, Adriano R. Lameira & Kirsten Bohn. 2014. Animal vocal sequences: Not the Markov chains we thought they were. *Proceedings of the Royal Society B: Biological Sciences* 281, 20141370. doi:10.1098/rspb.2014.1370
- Knowlton, Barbara J. & Larry R. Squire. 1996. Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning Memory and Cognition* 22, 169–181.
- Lai, Jun & Fenna H. Polietiek. 2011. The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition* 118, 265–273. doi:10.1016/j.cognition.2010.11.011
- Marcus, Gary F., S. Vijayan, S. B. Rao & Peter M. Vishton. 1999. Rule learning by seven-month-old infants. *Science* 283, 77–80. doi:10.1126/science.283.5398.77
- Moeser, Shannon D. & Albert S. Bregman. 1973. Imagery and language acquisition. *Journal of Verbal Learning and Verbal Behavior* 12, 91–98.
- Morgan, James L. & Elissa L. Newport. 1981. The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior* 20, 67–85.
- Mori, Kazuo & Shannon D. Moeser. 1983. The role of syntax markers and semantic referents in learning an artificial language. *Journal of Verbal Learning and Verbal Behavior* 22, 701–718.

- Mueller, Jutta L., Jörg Bahlmann & Angela D. Friederici. 2010. Learnability of embedded syntactic structures depends on prosodic cues. *Cognitive Science* 34, 338–349. doi:10.1111/j.1551-6709.2009.01093.x
- Mueller, Jutta L., Angela D. Friederici & Claudia Männel. 2012. Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences of the United States of America* 109, 15953–15958. doi:10.1073/pnas.1204319109
- Mueller, Jutta L., Angela D. Friederici & Claudia Männel. 2019. Developmental changes in automatic rule-learning mechanisms across early childhood. *Developmental Science* 22, e12700. doi:10.1111/desc.12700
- Muñoz, Carmen. 2006. Accuracy orders, rate of learning and age in morphological acquisition. In Carmen Muñoz (ed.), *Age and the Rate of Foreign Language Learning*, 107–126. Clevedon, United Kingdom: Multilingual Matters.
- Murphy, Robin A., Esther Mondragon & Victoria A. Murphy. 2008. Rule learning by rats. *Science* 319, 1849–1851. doi:10.1126/science.1151564
- Ojima, Shiro & Kazuo Okanoya. 2014. The non-hierarchical nature of the Chomsky hierarchy-driven artificial-grammar learning. *Biolinguistics* 8, 163–180.
- Oldfield, Richard C. 1971. Assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* 9, 97–113.
- Perruchet, Pierre & Arnaud Rey. 2005. Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review* 12, 307–313. doi:10.3758/BF03196377
- Pinker, Steven. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.
- Pinker, Steven. 1987. The bootstrapping problem in language acquisition. In Brian MacWhinney (ed.), *Mechanisms of Language Acquisition*, 399–441. Hillsdale, NJ: Lawrence Erlbaum.
- Reber, Arthur S. 1967. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 6:855–863.
- Reber, Arthur S. & Rhianon Allen. 1978. Analogic and abstraction strategies in synthetic grammar learning: Functionalist Interpretation. *Cognition* 6, 189–221.
- Rey, Arnaud, Pierre Perruchet & Joël Fagot. 2012. Centre-embedded structures are a by-product of associative learning and working memory constraints: Evidence from baboons (*Papio Papio*). *Cognition* 123, 180–184. doi:10.1016/j.cognition.2011.12.005
- Saffran, Jenny R., Richard N. Aslin, & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274, 1926–1928.
- Saffran, Jenny R. 2001. The use of predictive dependencies in language learning. *Journal of Memory and Language* 44, 493–515. doi:10.1006/jmla.2000.2759
- Saffran, Jenny R. 2002. Constraints on statistical language learning. *Journal of Memory and Language* 47, 172–196. doi:10.1006/jmla.2001.2839
- Savage-Rumbaugh, E. Sue, James L. Pate, Janet Lawson, S. Tom Smith & Steven Rosenbaum. 1983. Can a chimpanzee make a statement? *Journal of Experimental Psychology: General* 112, 457–492.



- Tomasello, Michael & Nameera Akhtar. 1995. Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development* 10, 201–224. doi:10.1016/0885-2014(95)90009-8
- Udden, Julia, Martin Ingvar, Peter Hagoort & Karl M. Petersson. 2012. Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: Investigating the push-down stack model. *Cognitive Science* 36, 1078–1101. doi:10.1111/j.1551-6709.2012.01235.x
- van Heijningen, Carolina A. A., Jos de Visser, Willem Zuidema & Carel ten Cate. 2009. Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences of the United States of America* 106, 20538–20543. doi:10.1073/pnas.0908113106
- Winkler, Marina, Jutta L. Mueller, Angela D. Friederici & Claudia Männel. 2018. Infant cognition includes the potentially human-unique ability to encode embedding. *Science Advances* 4, eaar8334. doi:10.1126/sciadv.aar8334

## Notice

Biolinguistics Editorial Team \*

\* Correspondence: [admin@biolinguistics.eu](mailto:admin@biolinguistics.eu)

It is easy to imagine what reflections on the year 2020 would look like. What makes it worse, though, is that for *Biolinguistics*—and in particular for me (Kleanthes)—the year could hardly have started worse. On 4 January 2020, Andreas Wadle passed away. Andy was my blood brother, one of my closest friends, for over 30 years. We first met at a fantasy and science fiction role-playing convention over 400 km from my hometown to which I hitch-hiked when I was 14 years old. Despite the distance, we became very close very quickly, and stayed very close even when distances grew further. We never lost touch, and while our mutual hobby interests may have faded over the years, we found other joint points of interest. Andy was also a gifted graphic designer. So it was Andy who first accompanied me on the design side when I started my CAT research group, the Cyprus Acquisition Team—and Andy also designed the logo for *Biolinguistics*, the original layout, and all artwork we had used over the years. *Andy, I love you.*

The year quickly brought about many changes that affected all of us. One change for *Biolinguistics* was the introduction of format-free initial submissions. That is, authors are now able to submit manuscripts without having to adhere to journal-specific formatting requirements. The only remaining requirements for initial submissions are that manuscripts include page and line numbers, are double spaced, and list references in a consistent manner. Another change was the new layout for published articles which now includes links to the authors' ORCIDs as well as an article's submission history. These changes are part of our efforts to streamline processes, presentation, and publication at *Biolinguistics*.

A small number of peer-reviewed articles were published this year, while others are still undergoing review and revision, and we expect a fuller volume 15 in 2021. Along the way, we achieved a milestone in 2020: Our 400<sup>th</sup> submission since the journal's founding. We are also completing the Special Issue on 'Biolinguistic Research in the 21<sup>st</sup> Century', guest-edited by Lluís Barceló-Coblijn and our Associate Editor Evelina Leivada. The first papers can already be downloaded as advance online publications from our website. Once complete, the entire special issue will be available this coming spring.

Lastly, behind the scenes the Editorial Team is also actively involved in efforts that seek to strengthen the growing worldwide efforts of platinum open

access journals such as *Biolinguistics* by actively cooperating in the efforts of the Free Journal Network (FJN). Hence, we are happy that our Associate Editor Patrick C. Trettenbrein has been elected to serve on the Board of Directors of FJN.

To continue our annual custom, we would like to express a big, heartfelt thank-you to all the reviewers who have served *Biolinguistics* throughout 2020, despite the many competing demands on their time in this tumultuous year. They are listed below by name (in alphabetical order). For everything else, we thank all our authors, readers, and supporters as well as the members of the *Biolinguistics* Advisory Board, the Editorial Board, and the Task Team for their participation and feedback, which is so crucial to the success of the journal. We appreciate your service to our scientific community.

### Reviewers

Daniel Adrover Roig  
Boban Arsenijevic  
Eran Asoulin  
Tej Bhatia  
Andreas Blümel  
Bart de Boer  
Derya Çokal  
Elizabeth Ellis  
Laura Garnham  
Martin Haspelmath  
Marc D. Hauser

D. Robert Ladd  
Hèctor M. Manrique  
Fernando Martínez Manrique  
Dimitrios Ntelitheos  
Anna Pineda Cirera  
Gareth Roberts  
Maggie Tallerman  
Arhonto Terzi  
Charalambos Themistocleous  
Constantina Theofanopoulou  
Veno Volenec

# *Biolinguistics*

Volume 14

2020

## TABLE OF CONTENTS

- |               |   |   |
|---------------|---|---|
| 001           | Attention to People Like You: A Proposal Regarding Neuroendocrine Effects on Linguistic Variation | Míša Hejná<br><i>Aarhus University</i><br>Lauren Ackerman<br><i>Newcastle University</i><br>Joel C. Wallenberg<br><i>Newcastle University</i> |
| 021           | Children's Learning of a Semantics-Free Artificial Grammar with Center Embedding                  | Shiro Ojima<br><i>University of Tokyo</i><br>Kazuo Okanoya<br><i>University of Tokyo</i>  |
| ★ FORUM ★ 049 | Notice  | <i>Biolinguistics</i> Editorial Team  |



biolinguistics

ISSN 1450-3417

*Biolinguistics* 14: 1-50, 2020  
<http://www.biolinguistics.eu>