Non-word repetition in children learning Yélî Dnye

Alejandrina Cristia[1] & Marisa Casillas[2,3]

[1] Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University

[2] Max Planck Institute for Psycholinguistics

[3] University of Chicago

Abstract

In non-word repetition (NWR) studies, participants are presented auditorily with an item that is phonologically legal but lexically meaningless in their language, and asked to repeat this item as closely as possible. NWR scores are thought to reflect some aspects of phonological development, saliently a perception-production loop supporting flexible production patterns. In this study, we report on NWR results among children (N = 40, aged 3–10 years) learning Yélî Dnye, an isolate spoken on Rossel Island in Papua New Guinea. Results make three contributions that are specific, and a fourth that is general. First, we found that non-word items containing typologically frequent sounds are repeated without changes more often that non-words containing typologically rare sounds, above and beyond any within-language frequency effects. Second, we documented rather weak effects of item length. Third, we found that age has a strong correlation with NWR scores, whereas there are weak correlations with child sex, maternal education, and birth order. Fourth, we weave our results with those of others to serve the general goal of reflecting on how NWR scores can be compared across participants, studies, languages, and populations, and the extent to which they shed light on the factors universally structuring variation in phonological development at a global and individual level.

Keywords: phonology, non-word repetition, Papuan, non-industrial, non-urban, comparative, typology, markedness, literacy

Word count: 11,500 words

Non-word repetition in children learning Yélî Dnye

Introduction

Children's perception and production of phonetic and phonological units continues developing well beyond the first year of life, even extending into middle childhood (e.g., Hazan & Barrett, 2000; Rumsey, 2017). Much of the evidence for later phonological development comes from non-word repetition (NWR) tasks. In the present study, we use NWR to investigate the phonological development of children learning Yélî Dnye, an isolate language spoken in Papua New Guinea (PNG), which has a large and unusually dense phonological inventory. This allows us to contribute data at the intersection of language typology, language acquisition, and individual variation, as presented in more detail below.

What is NWR?. In a basic NWR task, the participant listens to a production of a word-like form, such as /bilik/, and then repeats back what they heard without changing any phonological feature that is contrastive in the language. For instance, in English, a response of [bilig] or [pilik] would be scored as incorrect; a response [biːlik], where the vowel is lengthened without change of quality would be scored as correct, because English does not have contrastive vowel length.

NWR has been used to seek answers to a variety of theoretical questions, including what the links between phonology, working memory, and the lexicon are (Bowey, 2001), and how extensively phonological constraints found in the lexicon affect online production (Gallagher, 2014). NWR is also frequently used in applied contexts, notably as a diagnostic tool for language delays and disorders (Chiat, 2015; Estes, Evans, & Else-Quest, 2007). Since non-words can be generated in any language, it has attracted the attention of researchers working in multilingual and linguistically diverse environments, particularly in Europe in the context of diagnosing language impairments among bilingual children (Armon-Lotem, Jong, & Meir, 2015; Chiat, 2015; COST Action, 2009; Meir, Walters, & Armon-Lotem, 2016). NWR

tasks probably tap into many skills (for relevant discussion see Coady & Evans, 2008; Santos, Frau, Labrevoit, & Zebib, 2020). Non-words can be designed to try to isolate certain skills more narrowly; for instance, one can choose non-words that contain real morphemes in order to load more on prior language experience, or non-words that are shorter to avoid loading on working memory (see a discussion in Chiat, 2015). Broadly, however, NWR scores will necessarily reflect to a certain extent phonological knowledge (to perceive the item precisely despite not having heard it before) as well as online phonological working memory (to encode the item in the interval between hearing it and saying it back) and flexible production patterns (to produce the item precisely despite not having pronounced it before).

The present work. We aimed to contribute to four areas of research. We motivate each in turn.

NWR and typology.

The first research area is at the intersection of typology and phonological development. There has been an interest in adapting NWR to different languages, in part for applied purposes. In a review of NWR as a potential task to diagnose language impairments among bilingual children in Europe, Chiat (2015) discusses the impossibility of creating language-universal non-word items: Languages vary in their phonological inventory, sound sequencing (phonotactics), syllable structure, and word-level prosody. As a result, any one item created will be relatively easier if it more closely resembles real words in a language, making it difficult to balance difficulty when comparing children learning different languages. This previous literature also suggests some dimensions of difficulty—an issue to which we return in the next subsection.

Although this cross-linguistic literature is rich, the potential difficulty associated with specific phonetic targets composing the non-words has received relatively little attention. For example, Chiat (2015) discusses segmental complexity as a function of whether there are consonant clusters – which is arguably a factor reflecting phonotactics and syllable structure.

In the present study, we thought it was relevant to represent the rich phonological inventory found in Yélî Dnye, by including a variety of phonetic targets. Some of them are cross-linguistically rare, in that they are less common across languages than other sounds or phonetic targets. Phonologists, phoneticians, and psycholinguists have discussed the extent to which cross-linguistic frequency may reflect ease of processing and acquisition via diachronic language change. These works focus largely on phonotactics (Moreton & Pater, 2012) perceptual parsing of the (ambiguous) linguistic signal (Beddor, 2009; Ohala, 1981), and individual differences in processing styles (Bermúdez-Otero, 2015); small but significant effects that may cumulatively drive language change via phonologization (see Yu, 2021 for a recent review). Thus, the correlation between typological frequency and ease of acquisition is typically assumed to emerge from one or more of the following causal paths:

1. Sounds (and sound sequences) that are harder to perceive tend to be misperceived and thus lost diachronically
2. Sounds (and sound sequences) that are harder to pronounce tend to be mispronounced and thus lost diachronically
3. Sound sequences that are harder to hold in memory tend to be mispronounced and thus lost diachronically

Given these causal pathways, we predicted that variation in NWR across items will correlate with the cross-linguistic frequency of the phones composing those items.

Length effects on NWR.

The second research area we contribute data to is research looking at the impact of word length on NWR repetition within specific languages. Some work documents much lower NWR scores for longer, compared to shorter, items [e.g., among Cantonese-learning children; Stokes, Wong, Fletcher, and Leonard (2006)], whereas differences are negligible in other studies [e.g., among Italian learners; Piazzalunga, Previtali, Pozzoli, Scarponi, and Schindler (2019)].

It is possible that differences are due to language-specific characteristics, including the most common length of words in the lexicon and/or in child-experienced speech in that culture—a hypothesis discussed for instance in Chiat (2015) (pp. 7-8; see also p. 5). In broad terms, one may expect languages with a lexicon that is heavily biased towards monosyllables to show greater length effects than languages where words tend to be longer. A non-systematic meta-analysis does not provide overwhelming support for this hypothesis [Cristia and Casillas (2021); SM1].

Nonetheless, given the paucity of research looking at this question, and the diversity of current results, we did not approach this issue within a hypothesis-testing framework but sought instead to provide additional data on the question, which may be re-used in future meta- or mega-analytic analyses.

Individual variation correlations with NWR.

The third research area we contribute data to relates to the possibility that children differ from each other in NWR scores in systematic ways. Although the ideal systematic review is missing, a recent paper comes close with a rather extensive review of the literature looking at correlations between NWR scores and a variety of child-level variables, including familial socio-economic status, child vocabulary, and, among multilingual children, levels of exposure to the language on which the non-words are based (Farabolini, Rinaldi, Caselli, & Cristia, 2021). In a nutshell, most evidence is mixed, suggesting that correlations with individual variation may be small, and more data is needed to estimate their true size. For this reason, we descriptively report association strength between NWR scores and child age, sex, birth order, and maternal education.

Our focus on age stems from previous work, where performance increases with child age (Farmani et al., 2018; Kalnak, Peyrard-Janvid, Forssberg, & Sahlén, 2014; Vance, Stackhouse, & Wells, 2005). Although past research has not investigated potential correlations with birth

order on NWR, there is a sizable literature on these correlations in other language tasks (e.g., Havron et al., 2019), and therefore we report on these too. Common explanations for advantages for first- over later-born children include differential allocation of familial resources, particularly parental behaviors of cognitive stimulation (Lehmann, Nuevo-Chiquero, & Vidal-Fernandez, 2018). Regarding child sex, no significant correlation has been found in previous NWR research (Chiat & Roy, 2007), and in other language tasks evidence is mixed. Finally, prior research varies on whether significant differences as a function of maternal education are reported (Balladares, Marshall, & Griffiths, 2016; e.g., no difference found in Farmani et al., 2018; Kalnak, Peyrard-Janvid, Forssberg, & Sahlén, 2014; Meir & Armon-Lotem, 2017; Santos, Frau, Labrevoit, & Zebib, 2020; whereas differences were reported in Tuller et al., 2018). In other lines of work, maternal education correlates with child language outcomes, including vocabulary reports (Frank, Braginsky, Yurovsky, & Marchman, 2017) and word comprehension studies (Scaff, 2019). The causal pathways explaining this correlation are complex, but one explanation that is often discussed involves more educated mothers talking more to their children (see discussion in Cristia, Farabolini, Scaff, Havron, & Stieglitz, 2020).

NWR as a function of language and culture.

The fourth research goal we pursued is to use NWR with non-Western, non-urban populations, speaking a language with a moderate to large phonological inventory (see Maddieson, 2005 for a broad classification of languages based on inventory size). Indeed, NWR has seldom been used outside of urban settings in Europe and North America (Cristia, Farabolini, Scaff, Havron, & Stieglitz, 2020; with exceptions including Gallagher, 2014). To our knowledge, it has never been used with speakers of languages having large phonological inventories (e.g., more than 34 consonants and 7 vowel qualities Maddieson, 2013b, 2013a).

There are no theoretical reasons to presume that the technique will not generalize to these new conditions. That said, Cristia, Farabolini, Scaff, Havron, and Stieglitz (2020) recently reported relatively lower NWR scores among the Tsimane', a non-Western rural population,

interpreting these findings as consistent with the hypothesis that lower levels of infant-directed speech and/or low prevalence of literacy in a population could lead to population-level differences in NWR scores.

In view of these results, it is important to bear in mind that NWR is a task developed in countries where literacy is widespread, and it is considered an excellent predictor of reading; for instance, better than rhyme awareness (e.g., Gathercole, Willis, & Baddeley, 1991). Therefore, it may not be a general index of phonological development, but instead reflect certain non-universal language skills. Indeed, Cristia, Farabolini, Scaff, Havron, and Stieglitz (2020) present their task as being a good index of the development of "short-hand-like" representations specifically, which could thus miss, for example, more holistic phonological and phonetic representations. We return to the question of what was measured here in the Discussion.

Aside from Cristia, Farabolini, Scaff, Havron, and Stieglitz (2020)'s hypotheses just mentioned, we have found little discussion of linguistic differences (i.e., potential differences in NWR as a function of which specific language children are learning, and/or its typology) or cultural differences (i.e., potential differences in NWR as a function of other differences across human populations).[1]

[1]Please note that the linguistic and cultural differences discussed here are different from the differences discussed in the extensive literature on NWR by bilingual participants. In that literature, authors are concerned with individual variation in exposure to one (as opposed to other) languages among multilingual children, as variation in relative language experiences could mask potential effects of language impairment. To try to measure language abilities above and beyond relative levels of experience with a given language, authors have tried to build non-words that tap language-dependent or language-independent knowledge. For instance, Tuller et al. (2018) employed a set of non-words judged to be language independent and two others that were more aligned with either French or German. The intuition is that NWR will correlate with the relative levels of exposure to that language bilingual children more strongly when items are aligned with a specific language ("language-dependent") than when they are "language-independent." To make this more precise, among bilingual children, those that have more experience with English than Spanish should perform better on English non-words than their peers with less English experience. Preliminary results of an ongoing meta-analysis suggest significant associations between exposure to a given language and perfor-

Regarding potential language differences, we note that studies compose items by varying syllable structure and word length, while preferring relatively simple and universal phones (notably relying on point vowels, simple plosives, and fricatives that are prevalent across languages, like /s/). It would be interesting for future researchers to consider straying from the literature by varying other dimensions that are relevant to the language under study. For instance, for Yélî Dnye, it is relevant to vary phonological complexity of the individual sounds because of its large inventory.

Yélî Dnye phonology and community. Before going into the details of our study design, we first give an overview of Yélî Dnye phonology as well as a brief ethnographic review of the developmental environment on Rossel Island. As discussed above, NWR has been almost exclusively used in urban, industrialized populations, so we provide this additional ethnographic information to contextualize the adaptations we have made in running the task and collecting the data, compared to what is typical in commonly studied sites. Rossel Island lies 250 nautical miles off the coast of mainland PNG and is surrounded by a barrier reef. As a result, transport to and from the island is both infrequent and irregular. International phone calls and digital exchanges that require significant data transfer are typically not an option. Data collection is therefore typically limited to the duration of the researchers' on-island visits.

Yélî Dnye phonology.

Yélî Dnye is an isolate language (presumed Papuan) spoken by approximately 7,000 people residing on Rossel Island, an island found at the far end of the Louisiade Archipelago in Milne Bay Province, Papua New Guinea. The Yélî sound system, much like its baroque

---

mance in both language-dependent and language-independent NWR (Farabolini, Taboh, Ceravolo, & Guerra, 2021). In any case, this line of research focuses on links between exposure to a given language and NWR performance. In contrast, when we discuss linguistic or cultural differences here, we ask the question of whether children vary in their performance as a function of which language they are learning and/or their overall levels of language experience (not relative levels in a multilingual setting).

grammatical system (Levinson, 2021), is unlike any other in the region. In total, Yélî Dnye uses 90 distinctive segments (not including an additional three rarely used consonants), far outstripping the phoneme inventory size of other documented Papuan languages (Foley, 1986; Levinson, 2021; Maddieson & Levinson, in preparation). Thus, with respect to our first research goal, Yélî Dnye is a good language to use because its large phonological inventory includes sounds that vary in cross-linguistic frequency (including some rare sounds) that can be compared in the NWR setting.

To provide some qualitative information on this inventory, we add the following observations. With only four primary places of articulation (bilabial, alveolar, post-alveolar, and velar) and no voicing contrasts, the phonological inventory is remarkably packed with acoustically similar segments. The core oral stop system includes both singleton (/p/, /t/, /ʈ/[2], and /k/) and doubly-articulated (/tp/, /ʈp/, /kp/) segments, with a complete range of nasal equivalents (/m/, /n/, /ɳ/, /ŋ/, /nm/, /ɳm/, /ŋm/), and with a substantial portion of them contrastively pre-nasalized or nasally released (/mp/, /nt/, /ɳʈ/, /ŋk/, /nmtp/, /ɳmʈp/, /ŋmkp/, /ʈɳ/, /kŋ/, /ʈpɳm/, /kpŋm/). A large number of this combinatorial set can further be contrastively labialized, palatalized on release, or both (e.g., /pʲ/, /pʷ/, /pʲʷ/; /tpʲ/; /ɳmɖbʲ/; see Levinson (2021) for details). The consonantal inventory also includes a number of non-nasal continuants (/w/, /j/, /ɣ/, /l/, /βʲ/, /lʲ/, /lβʲ/). Vowels in Yélî Dnye may be oral or nasal, short or long. The 10 oral vowel qualities, which span four levels of vowel height, (/i/, /ɯ/, /u/, /e/, /o/, /ə/, /ɛ/, /ɔ/, /æ/, /ɑ/) can be produced as short and long vowels, with seven of these able to occur as short and long nasal vowels as well /ĩ/, /ũ/, /ə̃/, /ɛ̃/, /ɔ̃/, /æ̃/, /ɑ̃/).

Our second research goal is to measure the effect of non-word length on NWR, which may need to be interpreted taking into account typical word length in the language. We estimated

[2]We use Levinson's (2021) under-dot notation (e.g., /ʈ/) to denote the post-alveolar place of articulation; these stops are, articulatorily, somewhat variable in place, with at least some tokens produced fully sub-apically. In approximating cross-linguistic segment frequency below we use the corresponding retroflex for each stop segment (e.g., /ʈ/, /ʈp/, /ɳ/).

word length in words found in a conversational corpus (see Stimuli section for details), where the distribution of length was: 15% monosyllabic, 39% disyllabic, 29% trisyllabic, and the remaining 17% being longer than that. The vast majority of syllables use a CV format. A small portion of the lexicon features words with a final CVC syllable, but these are limited to codas of -/m/, -/p/, or -/j/ (e.g., ndap /ṇṭæp/ 'Spondylus shell') and are often resyllabified with an epenthetic /ɯ/ in spontaneous speech (e.g., ndapî /'ṇṭæpɯ/). There are also a handful of words starting with /æ/ (e.g., ala /æ'læ/ 'here') and a small collection of single-vowel grammatical morphemes (see Levinson (2021) for details).

Our knowledge of Yélî language development is growing (e.g., Brown, 2011, 2014; Brown & Casillas, in press; Casillas, Brown, & Levinson, 2021; Liszkowski, Brown, Callaghan, Takada, & de Vos, 2012), but research into Yélî phonological development has only just begun. For example, Peute and Casillas (In preparation) find that Yélî Dnye-learning children's early spontaneous consonant productions appear to exclusively feature simplex and typologically frequent phones. Other ongoing work on Yélî Dnye includes experiment-based infant phoneme discrimination data and errors made in elicited and spontaneous speech from young children, but these data are neither finalized nor yet externally reviewed (see Hellwig, Sarvasy, & Casillas, provisionally accepted for more information). These data will help better inform our current analyses based on NWR in the future (e.g., regarding common sound substitutions) but are not critical for testing our current question about the general correlation between cross-linguistic phone frequency and NWR performance.

Before closing this section, it bears mentioning that the language has an established orthography, which includes distinct graphemes for all the contrasts on which our items are based. Some children in our sample will have started school. Reading and writing instruction is currently done only in English (other than writing one's name). This was probably not the case for the majority of mothers of the children in our sample, who will have learned to read and write in Yélî Dnye during their first three years at school. It is possible that there is also some

home teaching of Yélî reading and writing, notably for reading the bible.

The Yélî community.

Some aspects of the community are relevant for contextualizing our study design and results, particularly regarding sources of individual variation. Specifically, we investigated potential correlations with age, child sex, maternal education, and birth order. There is nothing particular to note regarding age and child sex, but we have some comments that pertain to the other two factors.

The typical household in our dataset includes seven individuals (typically, a mixed-sex couple and children—their own and possibly some others staying with them, as discussed in the next paragraph) and is situated among a collection of four or more other households, with structures often arranged around an open grassy area. These household clusters are organized by patrilocal relation, such that they typically comprise a set of brothers, their wives and children, and their mother and father, with neighboring hamlets also typically related through the patriline. Land attribution for building one's home is decided collectively based on land availability.

Most Yélî parents are swidden horticulturalists, who occasionally fish. Within a group of households, it is often the case that older adolescents and adults spend their day tending to their farm plots (which may not be nearby), bringing up water from the river, washing clothes, preparing food, and engaging in other such activities. Starting around age two years, children more often spend large swaths of their day playing, swimming, and foraging for fruit, nuts, and shellfish in large (~10 members) independent and mixed-age child play groups (Brown & Casillas, in press; Casillas, Brown, & Levinson, 2021). Formal education is a priority for Yélî families, and many young parents have themselves pursued additional education beyond what is locally available (Casillas, Brown, & Levinson, 2021). Local schools are well out of walking distance for many children (i.e., more than 1 hour on foot or by canoe each day), so it is very common for households situated close to a school to host their school-aged relatives during the

weekdays for long segments of the school year. Children start school often at around age seven, although the precise age depends on the child's readiness, as judged by their teacher.

Some general ideas regarding potential correlations between our NWR measures and maternal education may be drawn from the observations above. To begin with, many of our participants above 6 years of age may not be living with their birth mother but with other relatives, which may weaken associations with maternal education. In addition, it seems to us that the length of formal education a given individual may have is not necessarily a good index of their socio-economic status or other individual properties, unlike what happens in industrialized sites, and variation may simply be due to random factors like living close to a school or having relatives there.

As for birth order, much of the work on correlations between birth order and cognitive development (including language) has been carried out in the last 70 years and in agrarian or industrialized settings (Barclay, 2015; Grätz, 2018), where nuclear families were more likely to be the prevalent rearing environment (Lancy, 2015). It is possible that birth order differences are stronger in such a setting, because much of the stimulation can only come from the parents. These effects may be much smaller in cultures where it is common for children to attend daycare at an early age (such as France) or where extended family typically live close by. The Yélî community falls in the latter case, as children are typically surrounded by siblings and cousins of several orders, regardless of their birth order in their nuclear family.

We add some observations that will help us integrate this study into the broader investigation of NWR across cultures. As mentioned previously, there is one report of relatively low NWR scores among the Tsimane', which the authors of that paper interpret as consistent with long-term effects of low levels of infant-directed speech (Cristia, Farabolini, Scaff, Havron, & Stieglitz, 2020). However, Cristia, Farabolini, Scaff, Havron, and Stieglitz (2020) also point out that this is based on between-paper comparisons, and thus methods and myriad other factors have not been controlled for. The Yélî community can help us gain new insights into this matter

because direct speech to children under 3;0 is comparably infrequent in this community (in fact it may be infrequent in many settings, including urban ones Bunce et al., under review), and additionally shares other societal characteristics wih the Tsimane' [e.g., is rural and relies on farming, children grow up in wide familial networks, etc; Casillas, Brown, and Levinson (2021)]. Although infant-directed speech has been measured in different ways among the Tsimane' and the Yélî communities, our most comparable estimates at present suggest that Tsimane' young children are spoken to about 4.2 minutes per hour (Scaff, Stieglitz, Casillas, & Cristia, 2021), and Yélî children about 3.6 minutes per hour (Casillas, Brown, & Levinson, 2021). Thus, if these input quantities in early childhood relate to lower NWR scores later in life, we should observe similarly low NWR scores here as in Cristia, Farabolini, Scaff, Havron, and Stieglitz (2020).

Research questions. After some preliminary analyses to set the stage, we perform statistical analyses to inform answers to the following questions:

- Does the cross-linguistic frequency of sounds in the stimuli predict NWR scores? Are rarer sounds more often substituted by commoner sounds?
- How do NWR scores change as a function of item length in number of syllables?
- Is individual variation in NWR scores attributable to child age, sex, birth order, and/or maternal education?

Throughout these analyses and in the Discussion, we also have in mind our fourth goal, namely integrating NWR results across samples varying in language and culture.

We had considered boosting the interpretational value of this evidence by announcing our analysis plans prior to conducting them. However, we realized that even pre-registering an analysis would be equivocal because we would not have enough power to look at all relationships of interest, in many cases possibly not enough to detect any of the known associations, given the previously discussed variability across studies. Therefore, all analyses in

the present study are descriptive and should be considered exploratory.

Methods

Participants. This study was approved as part of a larger research effort by the second author. The line of research was evaluated by the Radboud University Faculty of Social Sciences Ethics Committee (Ethiek Commissie van de faculteit der Sociale Wetenschappen; ECSW) in Nijmegen, The Netherlands (original request: ECSW2017-3001-474 Manko-Rowland; amendment: ECSW-2018-041), including the use of verbal (not written) consent. As discussed in subsection "The Yélî community," the combination of collective child guardianship practices and common hosting of school-aged children for them to attend school is that adult consent often comes from a combination of aunts, uncles, adult cousins, and grandparents standing in for the child's biological parents. Child assent is also culturally pertinent, as independence is encouraged and respected from toddlerhood (Brown & Casillas, in press). Participation was voluntary; children were invited to participate following indication of approval from an adult caregiver. Regardless of whether they completed the task, children were given a small snack as compensation. Children who showed initial interest but then decided not to participate were also given the snack.

We tested a total of 55 children from 38 families spread across four hamlet regions. We excluded test sessions from analysis for the following reasons: refused participation or failure to repeat items presented over headphones even after coaching ($N = 8$), spoke too softly to allow offline coding ($N = 5$), or were 13 years old or older ($N = 2$; we tested these teenagers to put younger children at ease). The remaining 40 children (14 girls) were aged from 3 to 10 years ($M = 6.40$ years, $SD = 1.50$ years). In terms of birth order, 6 were born first, 5 second, 2 third, 7 fourth, 5 fifth, and 1 sixth, with birth order missing for 14 children. These children were tested in a hamlet far from our research base, and we unfortunately did not ask about birth order before

leaving the site. Maternal years of education averaged 8.22 years (range 6-12 years).[3] We also note that there were 34 children only exposed to Yélî Dnye at home and 6 children exposed to Yélî Dnye plus one or more other languages at home.[4]

Stimuli. Many NWR studies are based on a fixed list of 12-16 items that vary in length between 1 and 4 syllables, often additionally varying syllable complexity and/or cluster presence and complexity, and always meeting the condition that they do not mean anything in the target language (e.g., Balladares, Marshall, & Griffiths, 2016; Wilsenach, 2013). We kept the same variation in item length and requirement for not being meaningful in the language, but we did not vary syllable complexity or clusters because these are vanishingly rare in Yélî Dnye. We also increased the number of items an individual child would be tested on, such that a child would get up to 23 items to repeat (other work has also used up to 24-46 items: Jaber-Awida, 2018; Kalnak, Peyrard-Janvid, Forssberg, & Sahlén, 2014; Piazzalunga, Previtali, Pozzoli, Scarponi, & Schindler, 2019), with the entire test inventory of 40 final items distributed across children. We used a relatively large number of items to explore correlations with length and phonological complexity. However, aware that this large item inventory might render the task longer and more tiresome, we split items across children. Naturally, designing the task in this way may make the study of individual variation within the population more difficult because different children are exposed to different items.

A first list of candidate items was generated during a trip to the island in 2018 by selecting simple consonants (/p/, /t/, /ṭ/, /k/, /m/, /n/, /w/, /y/) and vowels (/i/, /o/, /u/, /a/, /e/) and

[3]We asked for mothers' highest completed level of education. We then recorded the number of years entailed by having completed that level under ideal conditions.

[4]Most speakers of Yélî Dnye grow up speaking it monolingually until they begin attending school around the age of 7 years; school instruction is in English. While monolingual Yélî Dnye upbringing is common, multilingual families are not unusual, particularly in the region around the Catholic Mission (the same region in which much of the current data were collected), where there is a higher incidence of married-in mothers from other islands (Brown & Casillas, in press). Children in these multilingual families grow up speaking Yélî Dnye plus English, Tok Pisin, and/or other language(s) from the region.

combining them into consonant-vowel syllables, then sampling the space of resulting possible 2- to 4-syllable sequences. These candidates were automatically removed from consideration if they appeared in the most recent dictionary (Levinson, 2021). The second author presented them orally to three local research assistants, all native speakers of Yélî Dnye, who repeated each form as they would in an NWR task and additionally let the experimenter know if the item was in fact a word or phrase in Yélî Dnye. Any item reported to have a meaning or a strong association with another word form or meaning was excluded.

A second list of candidate items was generated in a second trip to the island in 2019, when data were collected, by selecting complex consonants and systematically crossing them with all the vowels in the Yélî Dnye inventory to produce consonant-vowel monosyllabic forms. As before, items were automatically excluded if they appeared in the dictionary. Additionally, since perceiving vowel length in isolated monosyllables is challenging, any item that had a short/long lexical neighbor was excluded. Additionally, we made sure that the precise consonant-vowel sequence occurred in some real word in the dictionary (i.e., that there was a longer word included the monosyllable as a sub-sequence). These candidates were then presented to one informant, for a final check that they did not mean anything. Together with the 2018 selection, they were recorded, based on their orthographic forms, using a Shure SM10A XLR dynamic headband microphone and an Olympus WS-832 stereo audio recorder (using an XLR to mini-jack adapter) by the same informant, monitored by the second author for clear production of the phonological target. The complete recorded list was finally presented to two more informants, who were able to repeat all the items and who confirmed there were no real words present. Despite these checks, one monosyllable was ultimately frequently identified as a real word in the resulting data (intended yî /yɯ/; identified as yi /yi/, ‘tree’). Additionally, an error was made when preparing files for annotation, resulting in two items being merged (tpâ /tpɑ/ and tp:a /tpæ̃/). These three problematic items are not described here, and removed from the analyses below.

The final list includes three practice items and 40 test items (across infants): 16 monosyllables containing sounds that are less frequent in the world's languages than singleton plosives; 8 bisyllables; 12 trisyllables; and 4 quadrisyllables (see Table ??).

Table 1

NWR stimuli in orthographic (Orth.) and phonological (Phon.) representations.

Practice		Monosyll		Bisyll		Trisyll		Tetrasyll	
Orth.	Phon.	Orth.	Phon.	Orth.	Phon.	Orth.	Phon.	Orth.	Phon.
nopimade	nɔpimæṭɛ	dp:a	ṭpã	kamo	kæmɔ	dimope	ṭimɔpɛ	dipońate	ṭipɔnætɛ
poni	pɔni	dpa	ṭpæ	kańi	kæni	diyeto	ṭijɛtɔ	ńomiwake	nɔmiwækɛ
wî	wɯ	dpâ	ṭpɑ	kipo	kipɔ	meyadi	mɛjæṭi	todiwuma	tɔṭiwumæ
		dpê	ṭpə	ńoki	nɔki	mituye	mitujɛ	wadikeńo	wæṭikɛnɔ
		dpéé	ṭpeː	ńomi	nɔmi	ńademo	næṭɛmɔ		
		dpi	ṭpi	piwa	piwæ	ńayeki	næjɛki		
		dpu	ṭpu	towi	tɔwi	ńuyedi	nujɛṭi		
		gh:ââ	ɣãː	tupa	tupæ	pedumi	pɛṭumi		
		ghuu	ɣuː			tiwuńe	tiwunɛ		
		kp:ââ	kpãː			tumowe	tumɔwɛ		
		kpu	kpu			widońe	wiṭɔnɛ		
		lv:ê	lβʲə̃			wumipo	wumipɔ		
		lva	lβʲæ						
		lvi	lβʲi						
		t:êê	tə̃ː						
		tpê	tpə						

A Praat script (Boersma & Weenink, 2020) was written to randomize this list 20 times, and to split it into two sub-lists, to generate 40 different elicitation sets. The 40 elicitation sets

are available online from osf.io/dtxue/. The split had the following constraints:

- The same three items were selected as practice items and used in all 40 elicitation sets.
- Splits were done within each length group from the 2018 items (i.e., separately for 2-, 3-, and 4-syllable items); and among onset groups for the difficult monosyllables generated in 2019 (i.e., all the monosyllables starting with /tp/ were split into 2 sub-lists). Since some of these groups had an odd number of items, one of the sub-lists was slightly longer than the other (20 vs. 23).
- Once the sub-list split had been done, items were randomized such that all children heard first the 3 practice items in a fixed order (1, 2, and 4 syllables), a randomized version of their sub-list selection of difficult onset items, and randomized versions of their 2-syllable, then 3-syllable, and finally 4-syllable items.

Cross-linguistic frequency.

To inform our analyses, we estimated the typological frequency of all phonological segments present in the target items using the PHOIBLE cross-linguistic phonological inventory database (Moran & McCloy, 2019). For each phone in our task, we extracted the number and percentage of languages noted to have that phone in its inventory. While PHOIBLE is a unprecedented in its scope, with phonological inventory data for over 2000 languages at the time of writing, it is of course still far from complete, which may mean that frequencies are estimates rather than precise descriptors. Note that nearly half of the phones in PHOIBLE are only attested in one language (Steven Moran, personal communication). Extrapolating from this observation, we treat the three segments in our stimuli that were unattested in PHOIBLE (/lβʲ/, /t̪p/, and /tp/) as having a frequency of 1 (i.e., appearing in one language), with a (rounded) percentile of 0% (i.e., its cross-linguistic percentile is zero).

Within-language frequency.

Additionally, we estimated the usage frequency of the phones present in the target items in

a corpus of child-centered recordings (Casillas, Brown, & Levinson, 2021). That corpus was constituted by sampling from audio-recordings (7–9 hours long), collected as 10 children aged between 1 month and 3 years went about their day. The researchers selected 9 2.5-minute clips randomly and 11 1- or 5-minute clips by hand (selected to represent peak turn-taking and child vocal activity). These clips were segmented and transcribed by the lead researcher and a highly knowledgeable local assistant, who speaks Yélî Dnye natively, has ample experience in this kind of research, and often knew all the recorded people personally. For more details, please refer to Casillas, Brown, and Levinson (2021).

For the present study, we extracted the transcriptions of adult speech (i.e., removing key child and other children's speech) and split them into words using white space. We then removed all English and Tok Pisin words. The resulting corpus contained a total of 18,934 word tokens of 1,686 unique word types. To get our phone frequency measure, we counted the number of word types in which the phone occurred, and applied the natural logarithm.[5] Here, unattested sounds were not considered (i.e., they were declared NA so that they do not count for analyses). Note that the resulting values estimate usage frequencies for very young children's input and, while this is somewhat different from what our older participants experience on a daily basis, we can expect that this is a reasonable approximation of the early input that formed the foundation of their phonological knowledge.

Procedure. There is some variation in procedure in previous work. For example, while items are often presented orally by the experimenter (Torrington Eaton, Newman, Ratner, & Rowe, 2015), an increasing number of studies have turned instead to playing back pre-recorded stimuli in order to increase control in stimulus presentation (Brandeker & Thordardottir, 2015).

In adapting the typical NWR procedure for our context, we balanced three desiderata:

[5] We also carried out analyses using token (rather than type) phone frequency, but this measure was not correlated with whole-item NWR scores, and therefore the fact that it did not explain away the predictive value of cross-linguistic phone frequency was less informative than the relationship discussed in the Results section.

That children would not be unduly exposed to the items before they themselves had to repeat them (i.e., from other children who had participated); that children would feel comfortable doing this task with us; and that community members would feel comfortable having their children do this task with us.

We tested in four different sites spread across the northeastern region of the island, making a single visit to each, conducting back-to-back testing of all eligible children present at the time of our visit in order to prevent the items from 'spreading' between children through hearsay. Whenever children living in the same household were tested, we tried to test children in age order, from oldest to youngest, to minimize intimidation for younger household members, and always using different elicitation sets. Because space availability was limited in different ways from hamlet to hamlet, the places where elicitation happened varied across testing sites. More information is available from the online supplementary materials.

We fitted the child with a headset microphone (Shure SM10A or WH20 XLR with a dynamic microphone on a headband, most children using the former) that fed into the left channel of a Tascam DR40x digital audio recorder. The headsets were designed for adult use and could not be comfortably seated on many children's heads without a more involved adjustment period. To minimize adjustment time, which was uncomfortable for some children given the proximity of the foreign experimenter and equipment, we placed the headband on children's shoulders in these cases, carefully adjusting the microphone's placement so that it was still close to the child's mouth. A research assistant who spoke Yélî Dnye natively, and who could also hear the instructions over headphones, sat next to the child throughout the task to provide instructions and, if needed, encouragement. The research assistant coached the child throughout the task to make sure that they understood what they were expected to do. Finally, an experimenter (the first author) was also fitted with headphones and a microphone; she was in charge of delivering the pre-recorded stimuli to the research assistant, the child, and herself over headphones.

The first phase of the experiment involved making sure the child understood the task. We explained the task and then orally presented the first practice item. At this point, many children did not say anything in response, which triggered the following procedure: First, the assistant insisted the child make a response. If the child still did not say anything, the assistant said a real word and then asked the child to repeat it, then another and another. If the child could repeat real words correctly, we provided the first training item over headphones again for children to repeat. Most children successfully started repeating the items at this point, but a few needed further help. In this case, the assistant modeled the behavior (i.e., the child and assistant would hear the item again, and the assistant would repeat it; then we would play the item again and ask the child to repeat it). A small minority of children still failed to repeat the item at this point. If so, we tried again with the second training item, at which point some children demonstrated task understanding and could continue. A fraction of the remaining children, however, failed to repeat this second training item, as well as the third one, in which case we stopped testing altogether (see Participants section for exclusions).

The second phase of the experiment involved going over the list of test items randomly assigned to each child. This was done in the same manner as the practice items: the stimulus was played over the headphones, and then the child repeated it aloud. NWR studies vary in whether children are allowed to hear and/or repeat the item more than one time. We had a fixed procedure for the test items (i.e., the non-practice items) in which the child was allowed to make further attempts if their first attempt was judged erroneous in some way by the assistant. The procedure worked as follows: When the child made an attempt, the assistant indicated to the experimenter whether the child's production was correct or not. If correct, the experimenter would whisper this note of correct repetition into a separate headset that fed into the right channel of the same Tascam recorder and we moved on to the next item. If not, the child was allowed to try again, with up to five attempts allowed before moving on to the next item. Children were not asked to make repetitions if they did not produce a first attempt. In total, test sessions took approximately six minutes, with the first minute attributed to practice and five

minutes to the actual test list.

Coding. The first author then annotated the onset and offset of all children's productions from the audio recording using Praat audio annotation software (Boersma & Weenink, 2020), then ran a script to extract these tokens, pairing them with their original auditory target stimulus, and writing these audio pairs out to .wav clips. The assistant then listened through all these paired target-repetition clips randomized across children and repetitions, grouped such that all the clips of the same target were listened to in succession. For each clip, the assistant indicated in a notebook whether the child production was a correct or incorrect repetition and orthographically transcribed the production, noting when the child uttered a recognizable word or phrase and adding the translation equivalent of that word/phrase into English. The assistant was also provided with some general examples of the types of errors children made without making specific reference to Yélî sounds or the items in the elicitation sets. Because the phonological inventory is so acoustically packed and because annotation was done based on audio data alone, it might be easy to mis-identify a segment, the assistant double-checked all of her annotations by listening to them and assessing them a second time, once she had completed a full first round.

Analyses. Previous work typically reports two scores: a binary word-level exact repetition score, and a phoneme-level score, defined as the number of phonemes that can be aligned across the target and attempt, divided by the number of phonemes of whichever item was longer (the target or the attempt; as in Cristia, Farabolini, Scaff, Havron, & Stieglitz, 2020). Previous work does not use distance metrics, but we report these rather than the phoneme-level scores because they are more informative. To illustrate these scores, recall our example of an English target being /bilik/ with an imagined response [bilig]. We would score this response as follows: at the whole item level this production would receive a score of zero (because the repetition is not exact); at the phoneme level this production would receive a score of 80% (4 out of 5 phonemes repeated exactly); and the phone-based Levenshtein distance for this

production is 20% (because 20% of phonemes were substituted or deleted). Notice that the phone-based Levenshtein distance is the complement of the phoneme-level NWR score. An advantage of using phone-based Levenshtein distance is that it is scored automatically with a script, and it can then easily be split in terms of deletions and substitutions (insertions were not attested in this study).

Results

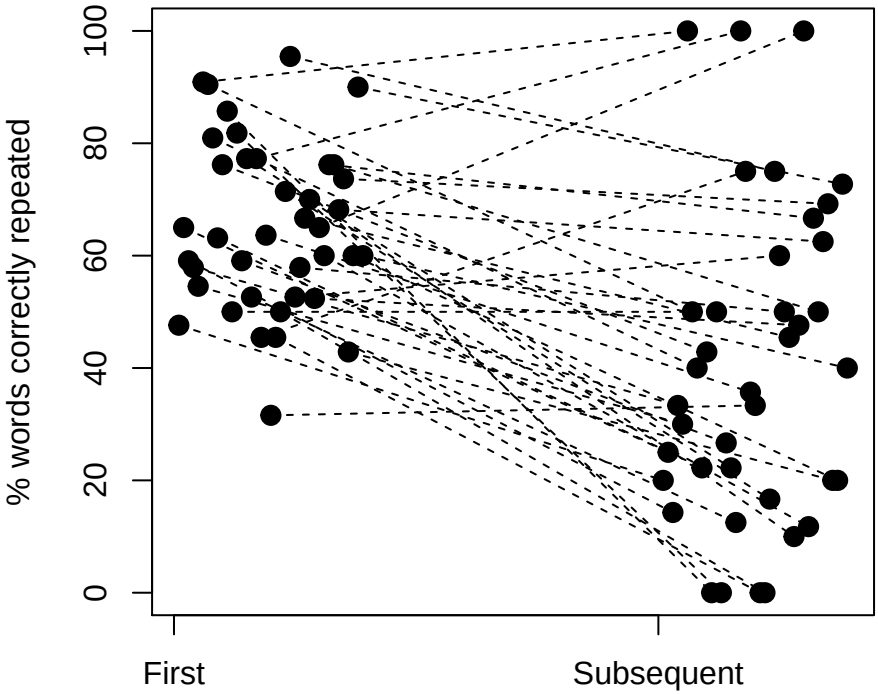

Figure 1. Whole-item NWR scores for individual participants averaging separately their first attempts and all other attempts.

Preliminary analyses. We first checked whether whole-item NWR scores varied between first and subsequent presentations of an item by averaging word-level scores at the participant level separately for first attempts and subsequent repetitions. We excluded 1 child who did not have data for one of these two types. As shown in Figure ??, participants' mean word-level scores became more heterogeneous in subsequent repetitions. Surprisingly, whole-item NWR scores for subsequent repetitions ($M = 40$, $SD = 28$) were on average lower than first ones (M

528 = 65, SD = 15), $t(38) = 5.89$, $p < 0.001$; Cohen's $d = 1.13$). Given uncertainty in whether
529 previous work used first or all repetitions, and given that score here declined and became more
530 heterogeneous in subsequent repetitions, we focus the remainder of our analyses only on first
531 repetitions, with the exception of qualitative analyses of substitutions.

532 Taking into account only the first attempts, we derived overall averages across all items.
533 The overall NWR score was M = 65% (SD = 15%), Cohen's $d = 4.39$. The phoneme-based
534 normalized Levenshtein distance was M = 21% (SD = 9%), meaning that about a fifth of
535 phonemes were substituted or deleted.

536 We also looked into the frequency with which mispronunciations resulted in real words.
537 In fact, two thirds of incorrect repetitions were recognizable as real words or phrases in Yélî
538 Dnye or English: 63%. This type of analysis is seldom reported. We could only find one
539 comparison point: Castro-Caldas, Petersson, Reis, Stone-Elander, and Ingvar (1998) found that
540 illiterate European Portuguese adults' NWR mispronunciations resulted in real words in 11.16%
541 of cases, whereas literate participants did so in only 1.71% of cases. The percentage we observe
542 here is much higher than reported in the study by Castro and colleagues, but we do not know
543 whether age, language, test structure, or some other factor explains this difference, such as the
544 particularities of the Yélî Dnye phonological inventory, which lead any error to result in many
545 true-word phonetic neighbors. Follow-up work exploring this type of error in children from
546 other populations in addition to further work on Yélî children may clarify this association.

547 NWR and typology: NWR as a function of cross-linguistic phone frequency. Turning to
548 our first research question, we analyzed variation in whole-item NWR scores as a function of the
549 average frequency with which sounds composing individual target words are found in languages
550 over the world. To look at this, we fit a mixed logistic regression in which the outcome variable
551 was whether the non-word was correctly repeated or not. The fixed effect of interest was the
552 average cross-linguistic phone frequency; we also included child age as a control fixed effect, in
553 interaction with cross-linguistic phone frequency, and allowed intercepts to vary over the