Open Library of Humanities

Proto-Lexicon Size and Phonotactic Knowledge are Linked in Non-Māori Speaking New Zealand Adults

Forrest Panther*, New Zealand Institute of Language, Brain and Behaviour, NZILBB, University of Canterbury, New Zealand, forrest.panther@canterbury.ac.nz

Wakayo Mattingley, New Zealand Institute of Language, Brain and Behaviour, NZILBB, University of Canterbury, New Zealand, wakayo.mattingley@canterbury.ac.nz

Simon Todd, Department of Linguistics, University of California, Santa Barbara, USA, sjtodd@ucsb.edu

Jennifer Hay, New Zealand Institute of Language, Brain and Behaviour, NZILBB, University of Canterbury, New Zealand; Department of Linguistics, University of Canterbury, New Zealand, jen.hay@canterbury.ac.nz

Jeanette King, New Zealand Institute of Language, Brain and Behaviour, NZILBB, University of Canterbury, New Zealand; Aotahi: School of Māori and Indigenous Studies, University of Canterbury, New Zealand, j.king@canterbury.ac.nz

*Corresponding author.

Most people in New Zealand are exposed to the Māori language on a regular basis, but do not speak it. It has recently been claimed that this exposure leads them to create a large proto-lexicon, consisting of implicit memories of words and word parts, without semantic knowledge. This yields sophisticated phonotactic knowledge (Oh, Todd, Beckner, Hay, King, & Needle, 2020). This claim was supported by two tasks in which Non-Māori-Speaking New Zealanders: (i) Distinguished real words from phonotactically matched non-words, suggesting lexical knowledge; (ii) Gave wellformedness ratings of non-words almost indistinguishable from those of fluent Māori speakers, demonstrating phonotactic knowledge.

Oh et al. (2020) ran these tasks on separate participants. While they hypothesised that phonotactic and lexical knowledge derived from the proto-lexicon, they did not establish a direct link between them. We replicate the two tasks, with improved stimuli, on the same set of participants. We find a statistically significant link between the tasks: Participants with a larger proto-lexicon (evidenced by performance in the Word Identification Task) show greater sensitivity to phonotactics in the Wellformedness Rating Task. This extends the previously reported results, increasing the evidence that exposure to a language you do not speak can lead to large-scale implicit knowledge about that language.

1 Introduction

In this paper, we present the results of a two-task experiment that explored the knowledge that Non-Māori Speaking New Zealanders (henceforth NMSs) have of Māori lexemes and phonotactics. This experiment extends the results of a previous study consisting of two experiments (Oh et al., 2020, described in section 1.3 below). The results of our experiment, as with the previous paper, show that NMSs have extensive knowledge of Māori words, and well-developed intuitions about Māori phonotactics. We go beyond the previous research in demonstrating a statistical association between individual participants' performance in these tasks, thus providing evidence in support of the hypothesis that these two forms of knowledge are causally linked.

First, we provide evidence in support of an adult 'proto-lexicon,' analogous to the proto-lexicon that infants develop for their first language. This proto-lexicon contains knowledge of wordforms that need not be associated with detailed morpho-syntactic or semantic content. The existence of this proto-lexicon among non-speakers of a language has important implications for understanding the development of multilingualism, particularly in terms of identifying the stages of development from monolingualism to bilingualism, and the extent to which latent long-term exposure results in comprehensive linguistic knowledge of a non-dominant language.

Second, we provide further supporting evidence in favour of the hypothesis that non-Māori speakers have extensive phonotactic knowledge. As with Oh et al. (2020), we show that non-Māori speakers are sensitive to phonotactics when rating the wellformedness of non-words.

Finally, we go beyond Oh et al. (2020) by demonstrating that the proto-lexion and phonotactic knowledge are linked in the same speakers. We interpret this as providing support for the assumption that the phonotactic knowledge is generated from the proto-lexicon. Participants with a larger proto-lexicon (as evidenced by their ability to distinguish real words from similiar non-words) have more sophisticated phonotactic knowledge, and are thus more sensitive to phonotactics when rating the wellformedness of non-words.

1.1 Language situation in New Zealand

Māori is the indigenous language of New Zealand. It is a Polynesian language in the Austronesian language family. In the 2018 New Zealand Census, 185,955 people self-reported as Māori speakers, representing about 3.9% of the total population of New Zealand (Statistics New Zealand, 2020).

Māori has a very small phoneme inventory, with ten consonant phonemes /p, t, k, m, n, ŋ, w, f, r, h/ and five vowels /i, e, a, o, u/. The five vowels also have long forms, which are indicated orthographically with a macron over the vowel. Because of this small inventory, when the Māori alphabet was developed by missionaries in the 1820s (Parkinson, 2016) the sound to spelling relationship was able to be fairly transparent, with the consonants represented by the letters <p, t, k, m, n, ng, w, wh, r, h>, and the vowels represented by <i, e, a, o, u>. Thus, written words

can be assumed to access a similar basis of knowledge representation as spoken words. This assumption is supported by psychological research that shows that reading involves phonological awareness (Ziegler & Goswami, 2005). In the experiments presented in this paper, we exploit this assumption by using written stimuli to probe phonological knowledge.

Māori vocabulary, as well as the use of Māori language, is pervasive in media, educational and cultural contexts in New Zealand. Māori expressions, including greetings, are largely and increasingly normalised. There are television stations and radio programs that broadcast in Māori, and in public buildings there is normally at least some Māori used or represented in writing. There are also public initiatives relating to the Māori language, including Māori Language Week. However, Māori is not a dominant language, and fluency is relatively rare. Consequently, people living in New Zealand receive consistent, low-level exposure to Māori, without ever learning to understand or speak it. The average non-Māori-speaking New Zealander is estimated to have an active Māori vocabulary of fewer than 100 words, as they can accurately match fewer than 100 words to their meaning in a multi-choice setting (Macalister, 2004).

This situation of a reasonable level of exposure without depth of vocabulary knowledge raises questions about the extent of the knowledge of Māori that NMSs have. How much phonological and lexical knowledge does this exposure result in? It is possible that the development of Māori knowledge from background exposure may reflect processes that occur the earliest stages of language acquisition. These are outlined in the next section.

1.2 Acquiring a lexicon

1.2.1 Statistical word segmentation

Literature on language acquisition has identified the importance of phonology in both first and second acquisition (Curtin & Hufnagle, 2009; Saffran, Werker, & Werner, 2007). The literature on the first language acquisition of phonology is primarily focused on the acquisition of phonemes (e.g., Kuhl, 2004; Wang, Seidl, & Cristia, 2021) and prosodic structure (Demuth, 2009). Other literature identifies the early acquisition of the ability to perceive recurring patterns in word forms (Chambers, Onishi, & Fisher, 2003; Coady & Aslin, 2004), even before children learn how to speak (Pierrehumbert, 2003). Before infants develop productive linguistic skills, they develop intricate speech perception abilities (Curtin & Archer, 2015). These perceptual skills underpin and aid in the development of later productive linguistic skills. Research has shown that by the age of six months, infants are already perceptually attuned to the phonology and prosody of their native language (Perszyk & Waxman, 2019; Johnson, Seidl, & Tyler, 2014; Shi, Werker, & Morgan, 1999; Shi, Morgan, & Allopenna, 1998). The effect of this is that even before infants begin to learn to use language, they have well-developed skills around the phonological structure of the language they are learning.

One key early step in acquiring a lexicon is identifying the boundaries between words. Infants younger than one use multiple cues to parse words in online speech, including word frequency (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005), position in the utterance (Johnson et al., 2014), syllable transitional probabilities (Saffran, Aslin, & Newport, 1996), segmental configuration (Archer & Curtin, 2016, 2011; MacKenzie, Curtin, & Graham, 2012), and probabilistic phonotactics (Archer & Curtin, 2016; Mattys & Jusczyk, 2001; Zamuner, 2009). On this latter point, infants use statistical information about the probabilities of segmental sequences to identify word boundaries. Statistical information about sound sequences is therefore important from the very early stages of language acquisition.

Laboratory experiments reveal that adults also use these statistical segmentation strategies to start to identify word boundaries in artifical languages that they are exposed to in an experimental setting (Peña, Bonatti, Nespor, & Mehler, 2002; Onnis, Monaghan, Richmond, & Chater, 2005; Newport & Aslin, 2000). Thus, it appears human listeners across the lifespan relatively automatically orient to a speech stream, and will unconsciously track statistical properties of that speech stream, in order to begin one of the earliest stages of language acquisition: Word segmentation.

1.2.2 The proto-lexicon

Once the speech segmentation process starts to identify candidate words, these words are stored in a *proto-lexicon.* A proto-lexicon is a mental lexicon consisting of sound sequences that are recognised and stored in long-term memory through language exposure, but it is not necessarily endowed with the semantic or morpho-syntactic properties associated with a fully-developed lexicon (Johnson, 2016; Ngon, Martin, Dupoux, Cabrol, Dutat, & Peperkamp, 2013, Martin, Peperkamp, & Dupoux, 2013; Hallé & de Boysson-Bardies, 1996). Evidence for the existence and nature of the infant proto-lexicon emerges from experimental research into the lexical knowledge of infants (see Junge, 2017; Swingley, 2009; Johnson, 2016 for reviews of this topic). Research has found that infants of six to nine months are able to recognise the meanings of certain high frequency words (Bergelson & Swingley, 2012, see Parise & Csibra, 2012, for similar results), but infants of around 11 months recognise word forms where they do not necessarily know the meaning of the word (Vihman, Nakai, DePaolis, & Hallé, 2004; see also Swingley, 2009, pp. 3619–3620, 3625–3626; Jusczyk, 2000; Swingley, 2005).

Importantly, the proto-lexicon is receptive, in that it is formed through language exposure, and is utilised in speech recognition but not yet in productive language use. Research emphasises the probabilistic nature of the proto-lexicon (Ngon et al., 2013): It is formed when infants parse speech streams into identifiable units, using the cues described in section 1.2.1. These units will often be words, but can also include highly frequent sound sequences that are not actually words in the language being learned (Ngon et al., 2013).

While most literature on the proto-lexicon focuses on infant language acquisition, there is also some evidence for proto-lexical knowledge in adults. For example in one language learning task, adults were exposed to an artificial language for 10 hours. An experiment conducted three years later revealed that participants could still recognise some high frequency words from that language (Frank, Tenenbaum, & Gibson, 2013). There is also evidence from former speakers of a language, who have lost their overt language knowledge, where they still show some evidence of implict knowledge of word-forms, thereby showing properties of a proto-lexicon (de Bot & Stoessel, 2000; van der Hoeven & de Bot, 2012; Hansen, Umeda, & McKinney, 2002). In fact, other research shows that non-speakers of Norwegian were able to distinguish words from non-words in Norwegian after a brief period of familiarisation (Kittleson, Aguilar, Tokerud, Plante, & Asbjørnsen, 2010). It seems likely, then, that adults who are exposed to a language in a real-life situation will start to automatically segment forms, and store them in a proto-lexicon.

1.2.3 Phonotactic knowledge and wellformedness

Related to, but separate from, the literature on word-segmentation is a literature within phonology, examining the degree to which phonological knowledge is gradient. This literature developed in response to the idea that phonology is based exclusively on categorical patterns that distinguish legal words from illegal words in a language. It shows that speakers of a language have fine-grained intuitions about degrees of wellformedness, showing that phonological knowledge is gradient, and is related to statistical patterns in the the language.

The main task used in the literature on phonotactic knowledge is ratings of non-words, and the overwhelming finding is that fluent speakers of a language provide ratings that are very highly correlated with lexical statistics calculated over word-types (not tokens) (Frisch, Large, Zawaydeh, & Pisoni, 2001; Hay, Pierrehumbert, & Beckman, 2004; Richtsmeier, 2011). In fact, approaches to phonotactics generally model the phonotactics of a language based on word-types in the lexicon of that language (Coleman & Pierrehumbert, 1997; Vitevitch & Luce, 1999; Frisch, Large, & Pisoni, 2000; Bailey & Hahn, 2001; Hay et al., 2004; Vitevitch & Luce, 2004). Thus, there is good evidence that once a lexicon is established, a speaker is able to generalise over the types in the lexicon in order to create strong intuitions about phonotactic wellformedness.

However, it is important to note that while word-segmentation strategies and phonotactic wellformedness intuitions are both probabilistic and related to each other, they draw on different types of specific knowledge. Information about phonotactic wellformedness cannot be straightforwardly extracted from a running speech stream if it relies on types in a lexicon. The statistics used for segmentation, by contrast, are tracked across tokens, and do not presuppose any knowledge of words (e.g., Cairns, Shillcock, Chater, & Levy, 1997). These 'bottom-up' approaches build phonotactic models directly from running speech, rather than word-types. It is important to note that while such approaches account for word segmentation, they do not

account for wellformedness judgments as effectively as models built over a lexicon. Indeed, Oh et al. (2020) compared this 'bottom-up' approach to a lexicon-based approach to modelling Māori phonotactics. They found that a phonotactic model based on lexical forms accounted for phonotactic judgments of Māori and non-Māori speakers better than a 'bottom-up' approach.

There is less work looking at phonotactic knowledge in children, although work from different tasks certainly suggests that word-based phonotactic knowledge is available and used. For example Edwards, Beckman, and Munson (2004) report that three to nine year old children produce non-words faster and more accurately when they are more phonotactically well-formed, noting that knowledge of word phonotactics is a key part of acquisition: "An increase in vocabulary size does not simply mean that the child knows more words, but also that the child is able to make more and more robust phonological generalizations" (p. 434). Similarly, Storkel and Rogers (2000) report that 10- and 13-year-old children can more easily learn high probability non-words than low probability non-words.

Even infants have been shown to be sensitive to phonotactics in ways that are suggestive of early stages of lexical acquisition. For example, infants show sensitivity to legal vs illegal words in their language from a young age (Steber & Rossi, 2020). Furthermore, there is evidence that sensitivity to gradient patterns exists as early as nine months: Jusczyk, Luce, and Charles-Luce (1994) found that nine month olds listened longer to high probability non-words than to ones containing low-probability phonotactic sequences.

Thus, while work on children has not specifically investigated the size and nature of the lexicon underpinning their phonotactic knowledge, the fact that it is evidenced so young would support the conjecture that explicit word knowledge is not a requirement. The word forms in the proto-lexicon seem to be sufficient to develop sophisticated probabilistic phonotactic knowledge.

In all, work on acquisition supports a trajectory in which learners use probabilistic information to segment words from the word stream, form a proto-lexicon, and then generate phonotactic knowledge from the lexicon. Of course this is not strictly sequential, and any knowledge in one of these domains can accelerate learning in the other.

Another relevant area of research is research into the phonotactic knowledge of bilinguals. Messer, Leseman, Boom, and Mayo (2010) compared the phonotactic knowledge of Turkish-dominant Turkish-Dutch bilingual preschoolers with Dutch monolingual preschoolers by conducting a wellformedness rating task with nonce words. They found that the bilingual participants had a stronger effect of phonotactics in Turkish in comparison with Dutch, while the monolingual participants outperformed the bilingual participants in Dutch. Similar results were found in research on infant and adult bilinguals in Catalan and Spanish, in which their response to phonotactically legal and illegal nonce words was affected by their dominant language (Sebastián-Gallés & Bosch, 2002). Indeed, research finds that even highly proficient L2 speakers

show transfer effects from L1, both in related language like English and German (Weber & Cutler, 2006), and very different languages like English and Cantonese (Yip, 2020). The effect of this is that listeners unconsciously use certain phonotactic cues from their L1 in order to segment speech in L2. Importantly, even if not to the same degree as monolinguals, the research generally supports phonotactic sensitivity in bilinguals to both their dominant and their non-dominant language (Frisch & Brea-Spahn, 2010).

Our work follows on from a previous study which considers the type of learning that may occur in adults who are exposed to a language regularly, but who are not overtly trying to learn that language. Do they learn to spot word boundaries and form a proto-lexicon? And what level of phonotactic knowledge might they generate?

1.3 Previous work

Oh et al. (2020) reported that non-Maori speaking adults living in New Zealand have extensive phonotactic knowledge of Māori, which stems from a surprisingly large proto-lexicon obtained through their environmental exposure to the language. This claim was on the basis of two online experiments:

1) **_Word Identification Task_**: Participants were presented with 150 pairs of words and non-words. They were asked to rate on a Likert scale how confident they were that each stimulus is a real word.

2) **_Wellformedness Rating Task_**: Participants were presented with 240–320 Māori-like non-words. They were asked to rate on a Likert scale how Māori-like each non-word was.

Along with non-Māori speaking New Zealanders, fluent Māori speakers and Americans with no knowledge of Māori also participated in the experiments.

The study reported the following findings:

1. In the Word Identification Task, across five frequency categories (from high frequency words to low frequency words), NMSs rated words more highly than non-words.

2. NMSs' ratings of stimuli in both the Word Identification Task and the Wellformedness Rating Task were positively correlated with the phonotactic wellformedness of the stimulus.

3. In the Wellformedness Rating Task, NMSs significantly outperformed Non-Māori speakers in the United States, and their results were nearly identical to the results of the Māori speaking participants.

4. The phonotactic knowledge of the NMSs is best modelled under the assumption that lexical items are decomposed into parts that occur with statistical regularity in the language, called

morphs. Moreover, NMSs' knowledge can be best modelled by a proto-lexicon consisting of approximately 1500 relatively common morphs.

Oh et al. (2020) conclude that there is good evidence that non-Māori speakers in New Zealand have a proto-lexicon of approximately 1500 morphs. This proto-lexicon is responsible for their ability to distinguish real words from non-words (under the assumption the real words are in the proto-lexicon), and also for their fine-grained phonotactic knowledge (which arises as a statistical generalization over the forms in the proto-lexicon).

The two experiments in Oh et al. (2020) were conducted on separate groups of participants. This means that the results of these experiments could not be explicitly connected. While Oh et al. (2020) hypothesise that the phonotactic knowledge stems from proto-lexical knowledge, they do not show a relationship between the two tasks. If there was a causal relationship, then participants with a greater ability to distinguish real words from non-words would also show increased phonotactic knowledge. The relationship between lexical and phonotactic knowledge is crucial to understanding the source of these effects, and consequently the inability to connect these experiments is a key limitation in the previous study. This paper conducts these two tasks on the same set of participants, allowing us to test whether there is a statistical link between responses in the two tasks.

In rerunning the experiments on a new set of participants, we also use an improved set of stimuli, and thus attempt to replicate the key findings of Oh et al. (2020) with this new set of stimuli that address some potential limitations:

Construction of Non-Words: In the Word Identification Task reported in Oh et al. (2020) there was an attempt to match real words with non-words of the same length and approximate phonotactic score. However the degree to which the non-words really 'match' is entirely dependent on the phonotactic model. If there is any way in which the real words actually systematically differ from the non-words, then participants could use this difference to perform the task, rather than word knowledge itself. In our experiment we use word pairs where each non-word is manually created to be maximally matched to its real-word pair. This enables us to confirm that the result replicates with an entirely different means of non-word construction.

Morphologically Complex Stimuli: The original stimuli in the Wellformedness Rating Task varied in length, and were manipulated to reflect different phonotactic probabilities. Potential morphological parses were not considered, but the analysis of Oh et al. (2020) suggests not only that the proto-lexicon contain morphs (rather than unanalysed morphologically-complex words), but also that the participants applied statistical knowledge to analyse some of the stimuli as morphologically complex. Our study uses shorter stimuli, which are carefully examined from a variety of perspectives, to minimise the likelihood of perceived morphological complexity (see the Supplementary Materials).

Orthographic Cues: The results of Oh et al. (2020) showed a small effect by the presence of a macron in the stimuli. Non-Māori speakers showed a tendency to rate stimuli containing macrons as more Māori-like. Macrons are a very salient feature of Māori orthography for NMSs, and consequently the presence of a macron impacted the results of the experiments. Our experiment does not include any stimuli with macrons.

Our experiments thus use a different and much cleaner set of stimuli to replicate the key results of Oh et al. (2020), and to seek a link between their two tasks. This would add considerable weight to the claim that a Māori proto-lexicon leads to phonotactic knowledge.

2 Methods

2.1 Experimental tasks

We replicated the online *Word Identification* and *Wellformedness Rating* tasks that were conducted by Oh et al. (2020), to verify the existence of the proto-lexicon and its relationship to phonological knowledge with updated stimulus sets. All experimental protocols in the online experiment were approved by the Human Research Ethics Committee at the University of Canterbury. This section contains a description of the generation of the experimental stimuli, factors used in the statistical analysis of the experimental results, experimental procedure, participant details, and the procedure of the statistical analyses. A more comprehensive description of the methodology of this experiment is in the Supplementary Materials.

2.2 Stimuli and materials

The full set of stimulus materials consists of 1042 items: 521 Māori word and Māori-like non-word pairs in total, which span 21 different phonotactic shapes. All stimuli were two to three syllables long, and none of them contained long vowels. This set provided the stimuli for both the Word Identification Task, and the Wellformedness Rating Task.

Unlike in Oh et al. (2020), non-words were manually obtained through minimal manipulation of real-word counterparts. We began by identifying candidates from all two to three syllable words consisting of at least three phonemes that are attested in the dictionary. Words that had an English lookalike, and words that potentially involved reduplication were removed from consideration. We then altered these words by manipulating up to three phonemes. A successful paired non-word meets the following criteria: (i) The word and non-word have the same number of syllables; (ii) The word and non-word have the same number of phonemes; (iii) The word and non-word have consonants and vowels in the same positions; (iv) The word and non-word have extremely similar phonotactic scores, as defined below. The number of phonemes ranges from three to six phonemes. Steps were taken to ensure that neither the real words or the non-words would be recognised as morphologically complex.

Each word in the stimuli was grouped into one of three *frequency* bins. A frequency score was generated for each word stimulus based on its raw count from the concatenation of the MAONZE corpus (King, Maclagan, Harlow, Keegan, & Watson, 2010) and the Māori Broadcast Corpus (MBC) (Boyce, 2006), totalling approximately 1.35M words of transcribed speech. For the purpose of this experiment, a non-word was grouped into the same frequency bin as its paired word. In all, there were 64 pairs in the high-frequency category (100+ per million; 135+ times in corpus), 101 pairs in the mid-frequency category (6-99 per million; 7-134 times in corpus), and 107 in the low-frequency category (1–5 per million; 1–6 times in corpus), for a total of 272 words. The stimuli included an additional 249 pairs that were unattested in the corpus.

A *phonotactic score* was calculated using length-normalised log-probabilities from trigram-based n-gram language models trained on morph types obtained from words in the Te Aka Dictionary using the SRI Language Modeling Toolkit (SRILM). We obtained a list of all headwords in Te Aka Dictionary (Moorfield, n.d.) as well as their inflected (i.e. passive voice) forms, excluding affixes and proper nouns. For the purposes of constructing the phonotactic model, we ignored all vowel length distinctions, because this sort of model showed the highest correlation with human ratings in the experiments reported by Oh et al. (2020).

Each stimulus in the dataset was assigned the number of phonological neighbours it has, in order to estimate the *neighbourhood density* of each stimulus. For the purpose of this study, we estimated this through a "neighbourhood occupancy rate" measure. This corresponds to a measure of the number of variations on a stimulus with an edit distance of one that correspond to a real Māori word.

For the Word Identification Task, the set of stimuli consisted of 272 word-nonword pairs for a total of 544 stimuli. 'Unattested' items (249 pairs) were not included in this task. The frequency values 'high' ($n_{pairs} = 64$), 'mid' ($n_{pairs} = 101$), and 'low' ($n_{pairs} = 107$) were used as bins. From this set of stimuli, for each participant, 20 pairs were randomly sampled without replacement from each bin, for a total of 120 stimuli. Each participant in the experiment received a different random sample. For each participant, no stimuli were shared between the Wellformedness Rating Task and the Word Identification Task.

For the Wellformedness Rating Task, the set of stimuli consisted of 521 Māori-like non-words. This dataset consisted of all non-words in the dataset. This includes non-words paired with unattested words (and thus excluded from the Word Identification Task). These non-words were grouped into three roughly equally-sized bins based on their phonotactic score: High ($n = 175$), medium ($n = 174$), low ($n = 172$). From this set, for each participant, 40 stimuli were sampled without replacement from each bin, for a total of 120 stimuli. Each participant in the experiment received a different random sample.

In order to investigate the correlation between performance in the Word Identification Task and Wellformedness Task, performance in the Word Identification Task was scored for each participant. We used this overall measure of participant accuracy as a predictor for our analysis of the Wellformedness Task. We estimate Word Identification Task accuracy for a participant by subtracting their mean confidence rating for non-words from their mean confidence rating for words:

$$Word\ Identification\ Task\ Score\ =\ mean(word\ rating)\ -\ mean(nonword\ rating)$$

Therefore, higher positive Word Identification Task scores indicate that the participant is more likely to reliably separate words from non-words, than participants who have lower scores.

2.2.1 Post-experiment questionnaire

A post-experiment questionnaire containing 26 questions was used in the experiment, which was almost identical to that of Oh et al. (2020). The questions were mostly about participants' sociolinguistic background information. In addition, the participants were asked about their levels of Māori proficiency and exposure as well as their knowledge of Māori. Because we wanted participants who had lived most of their life in New Zealand we added a question about whether they had lived overseas for more than one year and another question about whether they had studied linguistics at university. A couple of questions about participants' social attitude towards the Māori language were also added. The questions used in this questionnaire are in the Supplementary Materials.

2.3 Procedure

The experiment was conducted online using a custom in-browser interface (Chan, 2018). The experiment consisted of the Word Identification Task and Wellformedness Rating Task. Each task contained 120 trials, for a total of 240 trials. The order of these tasks was counterbalanced. Separate task instructions were presented prior to each task. The entire procedure took less than 30 minutes.

The procedure for each trial was as follows:

Word Identification Task: Participants were told to judge stimuli without looking up the word in a dictionary, or asking anyone else for help. A stimulus was presented orthographically in the middle of the screen, which was either a real word or a Māori-like non-word. Participants were asked to rate how confident they are that each item they saw was an actual Māori word, using a scale ranging from 1 ('Confident that it is NOT a Māori word') to 5 ('Confident that it IS a Māori word'). After the participant clicked one of the options and 'Next,' the next stimulus was presented.

Wellformedness Rating Task: Participants were instructed that they would see a made-up word in the middle of the screen. Participants were asked to rate how good this word would be as a Māori word, using a scale ranging from 1 ('Non Māori-like non-word') to 5 ('Highly Māori-like non-word'). Rating examples were given with a highly Māori-like non-word and non Māori-like non-word which were not actual stimulus items. After the participant clicked one of the options and 'Next,' the next stimulus was presented.

2.4 Participants

A total of 221 online participants were recruited using paid Facebook advertisements. Participants could choose to be paid a $10 online gift voucher. We set several criteria for participants results to be used for data analysis. These are summarised as follows:

- Be a native speaker of New Zealand English and 18 years or older.
- Not have lived outside New Zealand, for any period of longer than a year, since they were aged seven.
- Never have studied linguistics at a university.
- Not be able to hold a basic conversion in Māori.

Following screening on the above criteria, and further outlier removal (see the Supplementary Materials for details), there were 187 participants in total. Four participants were removed from analysis for the Word Identification Task and seven participants were removed from analysis of the Wellformedness Rating Task, due to low variation in the distribution of their responses.

A majority of participants were female (145, 77.5%) and lived on the North Island of New Zealand (131, 70.1%). They generally rated themselves very poorly for ability to speak or understand Māori. The survey included two questions that asked participants to rate themselves on a zero to five scale for their ability to speak Māori, and a zero to five scale for their ability to understand Māori, with zero indicating 'not at all,' and five indicating 'very well.' These measures were added up into a 10-point score to represent the Māori language ability of each participant. 149 (81.4%) participants scored at most two on this 10-point scale, and due to data filtering the maximum score was four, which 14 (7.7%) received. Although participants generally rated themselves as unable to speak or understand Māori, they indicated that they are exposed to Māori regularly. The survey also included two questions that asked participants to rate themselves on a zero to five scale for how often they were exposed to Māori through the media in everyday life, and how often they were exposed to it through social interactions. In this case, zero represented 'less than once a year' and five represented 'multiple times a day.' The two measures were again added up into a single 10-point score, this time representing the degree of

a participant's exposure to Māori in everyday life. 155 (82.9%) participants scored six or greater, representing exposure to Māori of some form on at least a monthly or weekly basis; thus, most participants generally received consistent exposure to Māori.

2.5 Statistical analyses

Consistent with Oh et al. (2020), we analysed the data with (logit) ordinal regression using the ordinal package (Christensen, 2019) in R (R Core Team, 2021). Using stepwise regression, models are compared to each other to see which fits the best, based on the Akaike Information Criterion (AIC). Random effects structure was kept as maximal as possible while maintaining the ability to successfully create the model (Barr, Levy, Scheepers, & Tily, 2013). For the Word Identification Task, the dependent variable was confidence rating on a Likert scale. For the Wellformedness Rating Task, the dependent variable was wellformedness rating on a Likert scale. The random effects were grouped by participant (corresponding to participant ID) and word.

As part of the stepwise regression approach, multiple effects were modelled initially, and non-significant effects were removed. For the Word Identification Task, the fixed effects considered were *phonotactic score*, *neighbourhood density*, *number of phonemes*, *stimulus type* (words, non-words) and *frequency bin* (high, mid, low). Phonotactic score and stimulus type were our test predictors, while phonological neighbourhood density, number of phonemes, and frequency bin were control predictors. Phonotactic score, neighbourhood density, and number of phonemes were centred. We started with a model containing five-way interactions. We then pruned the model by removing non-significant factors, the full details of which are in the Supplementary Materials.

For the Wellformedness Rating Task, the fixed effects considered were *phonotactic score*, *neighbour density*, and *Word Identification Task score*. All factors were centered. A similar stepwise regression approach was used for modelling this task as with the Word Identification Task.

3 Results
3.1 Word identification task

In order to factor the contrasting distribution of high frequency stimuli vs. mid & low frequency stimuli into the mixed effects analysis, we used Helmert contrasts to compare low vs. mid bins, and low & mid vs. high bins. In the model, the dependent variable was the confidence rating. Phonotactic score, neighbourhood density, number of phonemes, stimulus type and frequency bin were used as fixed effects. Phonotactic score, stimulus type, and frequency bin were entered into the model as a three-way interaction. Participant and stimulus were random effects, with the fixed effects as a random slope for participant. A summary of fixed effects in this model is in **Table 1**; for full details including threshold values, see the Supplementary Materials.

Effects	Estimate	Standard Error	z Value	p Value
Phonotactic Score (norm.)	3.23	0.88	3.69	<0.001*
Stimulus Type – Word	1.0	0.10	9.60	<0.001*
Frequency Bin (low vs. mid)	0.06	0.08	0.82	0.41
Frequency Bin (low/mid vs. high)	0.06	0.05	1.09	0.28
Number of Phonemes (norm.)	0.44	0.12	3.63	<0.001*
Neighbourhood Density (norm.)	5.82	1.47	3.96	<0.001*
Ph. Score (norm.) * Stim. Type – Wd.	–1.38	0.95	–1.46	0.14
Ph. Score (norm.) * Fq. Bin (low vs. mid)	0.07	0.75	0.08	0.93
Ph. Score (norm.) * Fq. Bin (low/mid vs. high)	0.35	0.51	0.70	0.49
Stim. Type – Wd. * Fq. Bin (low vs. mid)	0.18	0.11	1.63	0.1
Stim. Type – Wd. * Fq. Bin (low/mid vs. high)	0.52	0.08	6.88	<0.001*
Ph. Score (n.) * Stim. Type – Wd. * Fq. Bin (l. vs. m.)	–0.21	1.05	–0.2	0.84
Ph. Score (n.) * Stim. Type – Wd. * Fq. Bin (l./m. vs. h.)	–1.56	0.73	–2.14	0.04*

Table 1: Fixed effects of the ordinal regression model results for Word Identification Task. Asterisked *p* values are statistically significant. The estimates are rounded to two significant figures.

This model shows a significant effect of phonotactic score, stimulus type, number of phonemes, and neighbourhood density. It also shows a significant effect in the three-way interaction between phonotactic score, stimulus type, and frequency bin. **Figure 1** shows this three-way interaction. First, this plot shows that for all stimulus type and bin categories, apart from high frequency words, there is a positive relationship between predicted rating and phonotactic score. Second, in all frequency bins there is a higher predicted rating for words than non-words, with a positive correlation between the frequency bin and the difference in rating between words and non-words. Third, high frequency words appear to show a negative, rather than positive, relationship with phonotactic score. In relation to this latter point, the wide confidence band of the high frequency word category indicates that the negative slope is not likely to be significant. Instead, it indicates that phonotactic score is not a significant predictor for the confidence rating of high frequency words: participants were able to generally recognise the high frequency words used in this experiment, regardless of their wellformedness. This interaction is highly consistent

with the idea of a proto-lexicon and can be interpreted as follows. Since participants are exposed to high frequency words so often, they are extremely likely to have entries for such words in their proto-lexicons and may even be consciously aware of them, yielding high confidence scores regardless of phonotactics. In other categories (i.e. for nonwords, and for lower frequency words), participants may or may not have an entry for a given stimulus in their proto-lexicon, or may be unaware of it, leading them to use the wellformedness of the stimulus as a proxy for whether it is likely to be a word or not. In summary, as with Oh et al. (2020), we find that participants generally responded to the phonotactic probability of the stimuli, even though they were not explicitly tasked to in this task. In addition to showing sensitivity to phonotactics, they are also able to distinguish words from non-words. However, in Oh et al. (2020), the relationship between frequency and confidence rating was mixed. In these results, there is a much clearer effect.
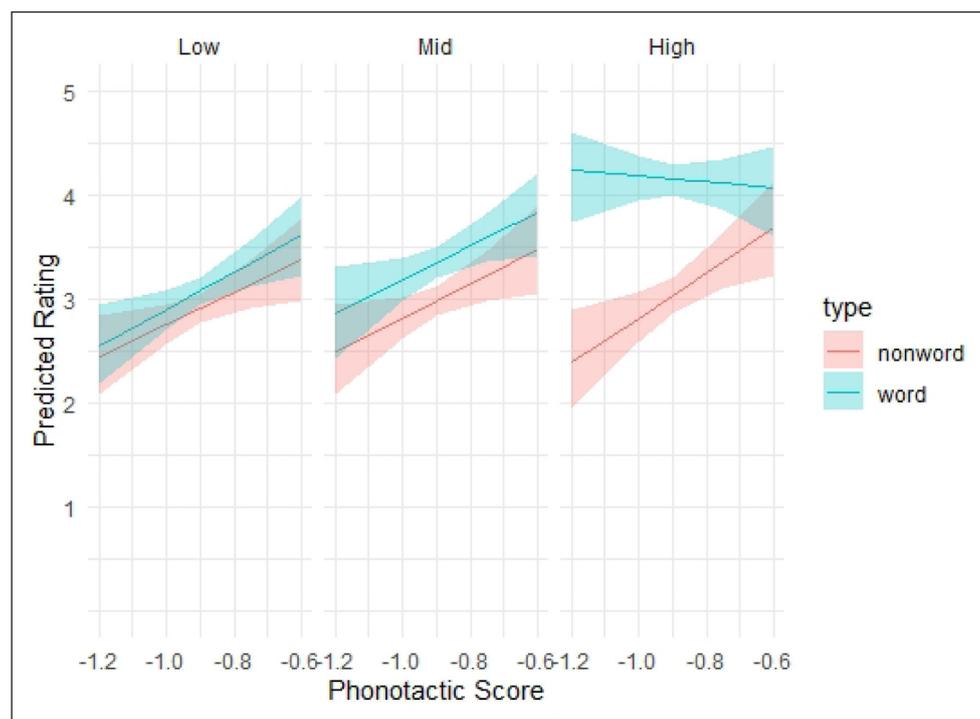

Figure 1: The relationship between predicted confidence rating and phonotactic score, stimulus type, and frequency bin in the Word Identification Task model.

The model also shows a significant positive effect of number of phonemes and neighbourhood density. Neighbourhood density is likely a predictor for a higher confidence rating because

stimuli with a denser phonological neighbourhood are more likely to resemble words known to the participants, and therefore be more easily confused with them. The positive effect of the number of phonemes is more difficult to interpret. We speculate that it relates to a possible effect of complexity in relation to length: stimuli that have more phonemes are more complex, and so may appear to be more word-like to participants if the participant is attempting to guess if a stimulus is a real word or not. This is only conjecture, however; we did not have a pre-existing hypothesis relating to this effect of number of phonemes. Consequently, we do not place much weight on this result.

3.2 Wellformedness Rating Task

The best model for the Wellformedness Rating Task is presented in **Table 2**. In the model, the dependent variable was the wellformedness rating. Phonotactic score, Word Identification Task score, and neighbourhood density were used as fixed effects. Phonotactic score was entered into the model with a two-way interaction with Word Identification Task score and neighbourhood density. Participant and stimulus were random effects, with phonotactic score + neighbourhood density, and Word Identification Task score as random slopes for these random effects respectively.

Effects	Estimate	Standard Error	z Value	p Value
Phonotactic Score (norm.)	6.84	0.467	14.6	<0.001*
Word Identification Task Score	−0.07	0.259	−0.3	0.793
Neighbourhood Density (norm.)	−1.24	0.830	−1.5	0.136
Phon. Score (norm.): Word Ident. Task Score	2.22	0.828	2.7	<0.01*
Phon. Score (norm.): Neigh. Density (norm.)	−17.77	7.666	−2.3	0.02*

Table 2: Fixed effects of the ordinal regression model results for Wellformedness Rating Task. Asterisked p values are statistically significant. All values are normalised. The estimates are rounded to two significant figures.

The model shows a strong effect of phonotactic score, and both interactions are significant: Phonotactic score and Word Identification Task Score, and phonotactic score and the neighbourhood density. **Figure 2** shows the mean rating for each stimulus in the Wellformedness Rating Task plotted against their phonotactic scores. The data shows that phonotactic score of stimuli corresponds positively with the wellformedness rating. The slope of the mean ratings is similar to the NMSs' ratings in Oh et al. (2020), and is steeper than the slope of the American ratings in that experiment. Consequently these results replicate Oh et al. (2020).
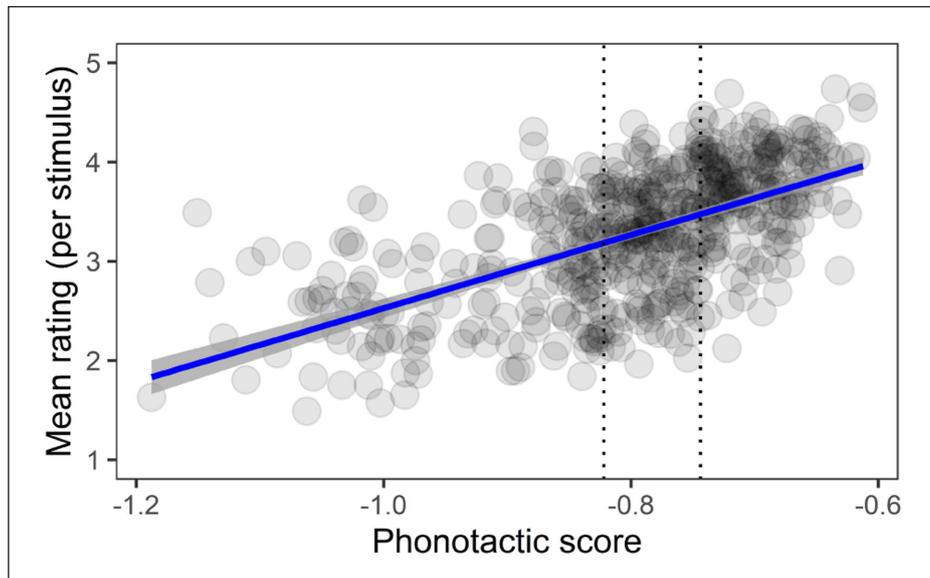
Figure 2: The relationship between phonotactic score and mean wellformedness rating of stimuli in the Wellformedness Rating Task in the raw experimental results. The dotted vertical bars indicate boundaries between phonotactic score bins (low-medium-high). The shaded area indicates the 95% confidence interval of the fitted line.

The phonotactic score of the stimuli interacts with two other factors in this model. Crucially, given our main research question, is that participant score in the Word Identification Task interacts with phonotactic score in predicting the wellformedness ratings. **Figure 3** shows this interaction from the model.
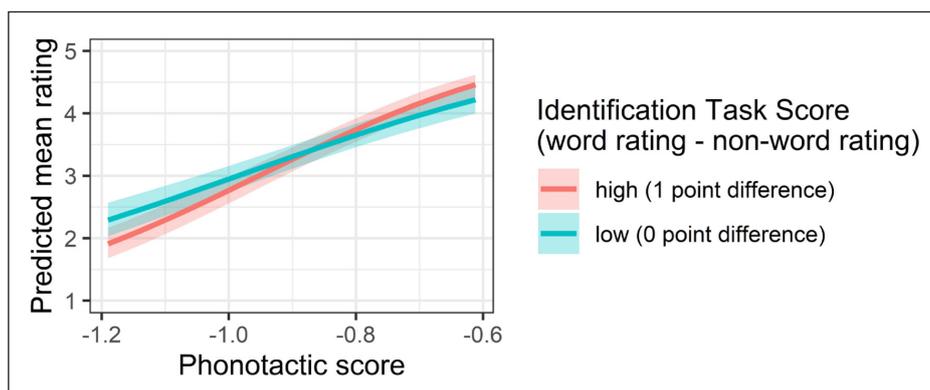
Figure 3: Performance of participants in the Wellformedness Rating Task based on their accuracy in the Word Identification Task (Word Identification Task score). The red line is ratings by participants with a Word Identification Task score of one. The teal line is ratings by participants with a Word Identification Task score of zero. The shaded areas indicate the 95% confidence interval for the predicted mean ratings.

Figure 3 shows the relationship between phonotactic score and mean rating for participants who performed well in the Word Identification Task (red) and those that performed less well (teal). The red line has a steeper slope, showing greater sensitivity to phonotactics. We can see that the better performing participants from the Wellformedness Rating Task (in red) generally gave more low ratings to forms with low phonotactic scores, and more high ratings to forms with high phonotactic scores. This significant interaction, then, shows the hypothesised link between the two tasks.

Finally, the second interaction involves the neighbourhood density of the stimulus. Oh et al. (2020) did not test for such an effect, but this was more important for our study, as our non-words were shorter, more closely modelled on real words, and thus more likely to have neighbours. The significant interaction indicates that non-words with a greater neighbourhood density do not benefit from higher phonotactic scores as strongly as non-words with a less dense neighbourhood. At first glance this appears somewhat counter intuitive, as two factors that should make a non-word seem more word-like combine to predict lower wellformedness ratings. We speculate that this relates to the nature of the task. Participants are told that the words are not words of Māori, and are asked to rate "how good the word would be as a Māori word." If it actually activates a real word, this may make it a poor candidate as a real word, as this could lead to potential lexical confusion. Thus, if the question was reworded to ask how much the word resembles existing words, this interaction may well disappear or reverse. This conjecture would need testing, and as this was not a predicted effect, we do not place too much weight on the result at this point.

The key results of the experiment are that: (a) Oh et al. (2020) is replicated with a new set of stimuli – non-Māori speakers remain highly sensitive to Māori phonotactics in this task, and (b) our research question is answered positively. Participants with a larger apparent proto-lexicon (as assessed through the Word Identification Task) are more sensitive to phonotactics (as assessed through the Wellformedness Rating Task).

4 Discussion

Using a completely different set of stimuli and participants, we replicate the key findings of Oh et al. (2020), namely that: (a) Non-Māori speakers in New Zealand can discriminate words from non-words; and (b) they are highly sensitive to Māori phonotactics. We provide an important demonstration that the key findings of Oh et al. (2020) replicate with a set of stimuli that is much more carefully controlled. In the previous study, for example, the morphological complexity of the stimuli was not controlled, because it only emerged as a relevant factor after data analysis began, but the conclusions make critical reference to morphological structure. Furthermore, the phonotactics of words and non-words were also not optimally matched in the previous work. A major contribution of this work is to demonstrate that, despite shortcomings in the stimuli of Oh et al. (2020), their major findings are robust.

In terms of the discrimination of words and non-words, the effect that we find is robust, but it strongly correlates with word frequency. In the high frequency bin, words are reliably discriminated by participants in a way that depends less on phonotactics. In the mid and low frequency bins, there is a stronger effect of phonotactics. This effect is likely to do with the reduced role of lexical knowledge at lower frequencies: Lower frequency words are less likely to be recognised by the participants, and consequently participants rely on probabilistic phonotactics to identify words at lower frequencies. Note that while this distribution differs from Oh et al.'s findings (2020, p. 3), it is consistent with their modelling procedures, which show that the phonotactic knowledge is best understood as stemming from a proto-lexicon of relatively frequent lexemes.

Crucially, our study conducted both tasks on the same set of participants. We were therefore able go significantly beyond Oh et al. (2020), by demonstrating a link between (proto)-lexical knowledge and phonotactic knowledge. Participants who can better distinguish real and non-words are also more sensitive to phonotactic probability in their ratings of a set of unrelated non-words.

We hypothesised that this would be so if the phonotactic knowledge was a direct consequence of the proto-lexicon, and the fact that we have found this correlation provides further support for that interpretation. Of course correlation is not complete proof of causation, and indeed, if the phonotactics was not causally linked to the proto-lexicon, it is not at all implausible to imagine that both were affected by a shared third factor, such as degree of exposure, or language learning aptitude. However the correlation between these tasks certainly adds to the weight of evidence that the phonotactic knowledge stems from lexical knowledge. Crucial here, too, is that fact that Oh et al. (2020) attempted multiple models of the phonotactic knowledge, including models that used sequences of phonemes as extracted from running speech, without knowledge of word boundaries. However the models that best explained the knowledge were over types in the lexicon – models that presuppose there is an inventory of words or morphs over which phonotactic generalizations can be made. We thus have multiple sources of converging evidence.

Our Word Identification Task shows that the ability of participants to identify monomorphemic words positively correlates with the frequency of the word. Oh et al's (2020) modelling shows the phonotactic knowledge is best explained by assuming knowledge of approximately 1500 common morphs, and our results show that speakers that can identify more morphs have more sophisticated phonotactic knowledge. This interpretation also retains a key assumption of the general literature on wellformedness rating tasks, namely that the phonotactic knowledge is a generalization over types in the lexicon (Frisch et al., 2001; Hay et al., 2004). Consequently, with evidence in the literature relating to the ability of individuals to segment words from the speech stream (Peña et al., 2002; Onnis et al., 2005; Newport & Aslin, 2000), the evidence points to a set of forms that have been segmented and stored in a proto-lexicon (Frank et al., 2013).

Our experiments support the interpretation that phonotactic knowledge appears to be derived from a proto-lexicon, but it is important to note that we do not make any further claims about the particular representation of that knowledge. It would be possible for such generalizations to be built on the fly, for example, via activation of words containing relevant sequences. It would also be possible for phonotactic knowledge to derive from generalizations in the lexicon, but for the lexical knowledge and phonotactic knowledge to be represented quite separately in the mind (Vitevitch & Luce, 1998, 1999). Some evidence for this account comes from reaction time data, showing that non-words tend to show facilitation effects from high phonotactic similiarity, whereas words tend to be slowed down by competition effects (Vitevitch & Luce, 1999). There is evidence of an effect of neighbourhood density in both the Word Identification Task and the Wellformedness Rating Task. In the Word Identification Task, neighbourhood density has an effect on ratings, but the modelling procedure did not show any significant interaction with words over nonwords, i.e. neighbourhood density had a positive effect on confidence ratings of stimuli generally, rather than just words. The Wellformedness Rating task shows a negative interaction between neighbourhood density and phonotactic score, which, as stated in section 3.2, we interpret in light of the nature of the task. Consequently our results do not support or refute this approach, given that we are not examining reaction times, and we did not design our stimuli to detect differences between phonotactic and neighbourhood effects; we included neighbourhood density as a control. Further experiments with non-Māori-speakers which are explicitly designed to probe questions relating to representation would be interesting. However the main conclusion of the current paper is that lexical knowledge feeds into phonotactic knowledge, and we do not claim anything further about how these types of knowledge are represented.

Our results, together with those reported in Oh et al. (2020), suggest that adult New Zealanders who do not speak Māori show the initial stages of language learning, similar to that identified in infants. The NMS's sensitivity to phonotactics calculated over lexical forms provides a direct analogue to work on first language acquisition (Jusczyk et al., 1994), providing further evidence that such phonotactic knowledge does not require overt word knowledge. Just as with work on infants (Ngon et al., 2013), these processes are unlikely to be strictly sequential, but as each type of knowledge increases it will feed back into other stages and accelerate further learning.

Even further than this, the performance of the participants in our research is comparable to that of bilinguals, especially in their wellformedness judgments. In section 1.2.3, we presented evidence in the literature that bilinguals are sensitive to the phonotactics of both of their languages, even if they show differences from monolinguals. In this experiment, we show that participants were highly sensitive to Māori phonotactics. While we did not compare fluent and non-fluent speaker judgments in this study, Oh et al. (2020) found that the judgments of non-Māori speakers were almost indistinguishable from that of fluent speakers. This result is significant, because it shows that the participants have highly-developed intuitions comparable

to that of bilinguals, despite not being obviously bilingual. The reason for this is undoubtedly related to long-term exposure to Māori.

One thing that we cannot know for certain, is how much exposure is needed to build up this kind of knowledge, nor whether adult exposure is sufficient, or whether exposure during childhood is also necessary. Our participants had all lived in New Zealand since they were seven-years-old, without any significant breaks, so this data-set does not allow us to disentangle this question. This is something we will be exploring in future work with participants with different types of exposure.

Nor can we be completely certain about the balance of different types of input in creating this knowledge. While we have focused on the fact that New Zealanders regularly hear spoken or sung Māori, it is also true that there is a reasonable amount of written Māori in the physical environment, in the form of street and town names, business names, email signatures, phrases in written reports, and so on. Unlike infants, our participants have an extra source of information in that they are literate. As written Māori contains word breaks (though not morpheme breaks), this may have somewhat contributed to identification of certain words. However, we consider it very unlikely that orthographical exposure is the key driver of our results, as New Zealanders who don't speak Māori are unlikely to visually attend to substantial excerpts of written Māori, whereas they can't help being exposed to substantial excerpts of spoken or sung Māori.

Significant further understanding of this phenomenon would come from work on exposure at different life stages, untangling the role of different types of exposure, and work on implicit learning of languages with less transparent orthography, and with different degrees of morphological transparency and complexity.

5 Conclusion

Oh et al. (2020) reported that non-Māori-speaking adult New Zealanders had a Māori proto-lexicon, and phonotactic sensitivity very similar to fluent speakers of Māori. Our work uses improved stimuli and a new set of participants, and draws the same conclusion. We go beyond Oh et al. (2020) by demonstrating a statistical link between their two experimental tasks. Participants who are best able to distinguish real Māori words from closely matched non-words are more sensitive to phonotactics in the Wellformedness Rating Task. This reinforces the interpretation that phonotactic knowledge emerges as a generalization over types in a lexicon – the more types you have, the more sophisticated the knowledge can be. It also suggests that this lexicon need not be overt: A proto-lexicon – a set of stored forms, which may not be associated with semantic knowledge or even overt awareness – can form a basis for phonotactic knowledge.

Supplementary materials

The supplementary materials of this paper are found at: https://github.com/FPanther/Replication22

Acknowledgements

Competing interests

The authors have no competing interests to declare.

References

Archer, S. L., & Curtin, S. (2011). Perceiving onset clusters in infancy. *Infant behavior and development, 34*(4), 534–540. DOI: https://doi.org/10.1016/j.infbeh.2011.07.001

Archer, S. L., & Curtin, S. (2016). Nine-month-olds use frequency of onset clusters to segment novel words. *Journal of experimental child psychology, 148*, 131–141. DOI: https://doi.org/10.1016/j.jecp.2016.04.004

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44*(4), 568–591. DOI: https://doi.org/10.1006/jmla.2000.2756

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253– 3258. DOI: https://doi.org/10.1073/pnas.1113380109

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological science, 16*(4), 298–304. DOI: https://doi.org/10.1111/j.0956-7976.2005.01531.x

Boyce, M. T. (2006). *A corpus of modern spoken Māori.* (Unpublished doctoral dissertation). Victoria University of Wellington.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*(2), 111–153. DOI: https://doi.org/10.1006/cogp.1997.0649

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77. DOI: https://doi.org/10.1016/s0010-0277(02)00233-0

Chan, C. L. (2018). *Speech In Noise 2.* Northwestern University. (Computer program)

Christensen, R. H. B. (2019). *ordinal—regression models for ordinal data.* (R package version 2019.12-10. https://CRAN.R-project.org/package = ordinal)

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of experimental child psychology*, *89*(3), 183–213. DOI: https://doi.org/10.1016/j.jecp.2004.07.004

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational phonology: Third meeting of the acl special interest group in computational phonology.* Retrieved from https://aclanthology.org/W97-1107

Curtin, S., & Archer, S. L. (2015). Speech perception. In E. L. Bavin & L. R. Naigles (Eds.), *The Cambridge Handbook of Child Language* (2nd ed., p. 137–158). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781316095829.007

Curtin, S., & Hufnagle, D. (2009). Speech perception. In E. Bavin (Ed.), *The Cambridge Handbook of Child Language* (1st ed., pp. 107–127). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781316095829.007

de Bot, K., & Stoessel, S. (2000). In search of yesterday's words: Reactivating a long-forgotten language. *Applied Linguistics*, *21*(3), 333–353. DOI: https://doi.org/10.1093/applin/21.3.333

Demuth, K. (2009). The prosody of syllables, words and morphemes. In *Cambridge Handbook of Child Language* (1st ed., pp. 183–98). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511576164.011

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. DOI: https://doi.org/10.1044/1092-4388(2004/034)

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (2013, 01). Learning and long-term retention of large-scale artificial languages. *PLOS ONE*, *8*, 1-6. DOI: https://doi.org/10.1371/journal.pone.0052500

Frisch, S. A., & Brea-Spahn, M. R. (2010). Metalinguistic judgments of phonotactics by monolinguals and bilinguals. *Laboratory Phonology*, *1*(2), 345–360. DOI: https://doi.org/10.1515/labphon.2010.018

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language*, *42*(4), 481–496. DOI: https://doi.org/10.1006/jmla.1999.2692

Frisch, S. A., Large, N. R., Zawaydeh, B., & Pisoni, D. B. (2001). Emergent phonotactic generalizations in english and arabic. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure.* Amsterdam/Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/tsl.45.09fri

Hallé, P. A., & de Boysson-Bardies, B. (1996). The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development, 19*(4), 463–481. DOI: https://doi.org/10.1016/S0163-6383(96)90007-7

Hansen, L., Umeda, Y., & McKinney, M. (2002). Savings in the relearning of second language vocabulary: The effects of time and proficiency. *Language Learning, 52*(4), 653–678. DOI: https://doi.org/10.1111/1467-9922.00200

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology vi.* Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511486425.004

Johnson, E. K. (2016). Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics, 2*, 391–412. DOI: https://doi.org/10.1146/annurev-linguistics-011415-040616

Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PloS one, 9*(1), e83546. DOI: https://doi.org/10.1371/journal.pone.0083546

Junge, C. (2017). The proto-lexicon: segmenting word-like units from speech. In G. Westermann & N. Mani (Eds.), *Early word learning* (pp. 15–29). Routledge. DOI: https://doi.org/10.4324/9781315730974-2

Jusczyk, P. W. (2000). *The Discovery of Spoken Language.* Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/2447.001.0001

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*(5), 630–645. DOI: https://doi.org/10.1006/jmla.1994.1030

King, J., Maclagan, M., Harlow, R., Keegan, P., & Watson, C. (2010). The MAONZE corpus: Establishing a corpus of Māori speech. *New Zealand Studies in Applied Linguistics, 16*(2), 1–16.

Kittleson, M. M., Aguilar, J. M., Tokerud, G. L., Plante, E., & Asbjørnsen, A. E. (2010). Implicit language learning: Adults' ability to segment words in norwegian. *Bilingualism: Language and Cognition, 13*(4), 513–523. DOI: https://doi.org/10.1017/S1366728910000039

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience, 5*(11), 831–843. DOI: https://doi.org/10.1038/nrn1533

Macalister, J. (2004). A survey of Māori word knowledge. *English in Aotearoa, 52*, 69–73.

MacKenzie, H., Curtin, S., & Graham, S. A. (2012). 12-month-olds' phonotactic knowledge guides their word–object mappings. *Child development, 83*(4), 1129–1136. DOI: https://doi.org/10.1111/j.1467-8624.2012.01764.x

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive science, 37*(1), 103–124. DOI: https://doi.org/10.1111/j.1551-6709.2012.01267.x

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*(2), 91–121. DOI: https://doi.org/10.1016/S0010-0277(00)00109-8

Messer, M. H., Leseman, P. P., Boom, J., & Mayo, A. Y. (2010). Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology, 105*(4), 306–323. DOI: https://doi.org/10.1016/j.jecp.2009.12.006

Moorfield, J. (n.d.). *Te Aka Online Māori Dictionary.* Retrieved from https://maoridictionary.co.nz/

Newport, E. L., & Aslin, R. N. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In *Proceedings of the 24th annual boston university conference on language development* (Vol. 1, pp. 1–21).

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (non) words,(non) words,(non) words: evidence for a protolexicon during the first year of life. *Developmental Science, 16*(1), 24–34. DOI: https://doi.org/10.1111/j.1467-7687.2012.01189.x

Oh, Y., Todd, S., Beckner, C., Hay, J., King, J., & Needle, J. (2020). Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific reports, 10*(1), 1–9. DOI: https://doi.org/10.1038/s41598-020-78810-4

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language, 53*(2), 225–237. DOI: https://doi.org/10.1016/j.jml.2005.02.011

Parise, E., & Csibra, G. (2012). Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. *Psychological Science, 23*(7), 728–733. DOI: https://doi.org/10.1177/0956797612438734

Parkinson, P. (2016). The Māori grammars and vocabularies of Thomas Kendall and John Gare Butler. *Asia-Pacific Linguistics, 26*, 1–163. Retrieved from http://hdl.handle.net/1885/104299

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*(5593), 604–607. DOI: https://doi.org/10.1126/science.1072901

Perszyk, D. R., & Waxman, S. R. (2019). Infants' advances in speech perception shape their earliest links between language and cognition. *Scientific reports, 9*(1), 1–6. DOI: https://doi.org/10.1038/s41598-019-39511-9

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech, 46*(2-3), 115–154. DOI: https://doi.org/10.1177/00238309030460020501

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Richtsmeier, P. T. (2011). Word-types not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology, 2*, 157–183. DOI: https://doi.org/10.1515/labphon.2011.005

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. DOI: https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Werker, J. F., & Werner, L. A. (2007). The infant's auditory world: Hearing, speech, and the beginnings of language. *Handbook of child psychology, 2*. DOI: https://doi.org/10.1002/9780470147658.chpsy0202

Sebastián-Gallés, N., & Bosch, L. (2002). Building phonotactic knowledge in bilinguals: role of early exposure. *Journal of Experimental Psychology: Human Perception and Performance, 28*(4), 974. DOI: https://doi.org/10.1037/0096-1523.28.4.974

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of child language, 25*(1), 169–201. DOI: https://doi.org/10.1017/S0305000997003395

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*(2), B11–B21. DOI: https://doi.org/10.1016/S0010-0277(99)00047-5

Statistics New Zealand. (2020). *2018 census totals by topic – national highlights.* Retrieved 2021-06-24, from https://www.stats.govt.nz/information-releases/2018-census-totals-by-topic-national-highlights-updated

Steber, S., & Rossi, S. (2020). So young, yet so mature? electrophysiological and vascular correlates of phonotactic processing in 18-month-olds. *Developmental cognitive neuroscience, 43*, 100784. DOI: https://doi.org/10.1016/j.dcn.2020.100784

Storkel, H. L., & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *clinical linguistics & phonetics, 14*(6), 407–425. DOI: https://doi.org/10.1080/026992000415859

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology, 50*(1), 86–132. DOI: https://doi.org/10.1016/j.cogpsych.2004.06.001

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3617–3632. DOI: https://doi.org/10.1098/rstb.2009.0107

van der Hoeven, N., & de Bot, K. (2012). Relearning in the elderly: age-related effects on the size of savings. *Language Learning, 62*(1), 42–67. DOI: https://doi.org/10.1111/j.1467-9922.2011.00689.x

Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language, 50*(3), 336–353. DOI: https://doi.org/10.1016/j.jml.2003.11.004

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in spoken word recognition. *Psychological Science*, 9–325. DOI: https://doi.org/10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of memory and language, 40*(3), 374–408. DOI: https://doi.org/10.1006/jmla.1998.2618

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers, 36*(3), 481–487. DOI: https://doi.org/10.3758/BF03195594

Wang, Y., Seidl, A., & Cristia, A. (2021). Infant speech perception and cognitive skills as predictors of later vocabulary. *Infant Behavior and Development, 62*, 101524. DOI: https://doi.org/10.1016/j.infbeh.2020.101524

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *The Journal of the Acoustical Society of America, 119*(1), 597–607. DOI: https://doi.org/10.1121/1.2141003

Yip, M. C. (2020). Spoken word recognition of l2 using probabilistic phonotactics in l1: evidence from cantonese-english bilinguals. *Language Sciences, 80*, 101287. DOI: https://doi.org/10.1016/j.langsci.2020.101287

Zamuner, T. S. (2009). Phonotactic probabilities at the onset of language development: Speech production and word position. *J Speech Lang Hear Res, 52*(1), 49–60. DOI: https://doi.org/10.1044/1092-4388(2008/07-0138)

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin, 131*(1), 3. DOI: https://doi.org/10.1037/0033-2909.131.1.3