UNIVERSITY OF CALIFORNIA

Santa Barbara

Exploring the Extent of Statistical Learning used by Implicit Language Learners: Insights from Non-Māori Speakers Exposed to Māori

A Thesis submitted in partial satisfaction of the
requirements for the degree Master of Arts
in Linguistics

by

Ashvini Varatharaj

Committee in charge:
Professor Simon Todd, Chair
Professor Laurel Brehm
Professor William Wang

March 2024

The thesis of Ashvini Varatharaj is approved.

____

Laurel Brehm

____

William Wang

____

Simon Todd, Committee Chair

March 2024

Exploring the Extent of Statistical Learning used by Implicit Language Learners: Insights from Non-Māori Speakers Exposed to Māori

by

Ashvini Varatharaj

ACKNOWLEDGEMENTS

I would like to express gratitude to my advisor, Professor Simon Todd, for his guidance, support, and valuable time throughout this project. His expertise, feedback and moral support during our weekly meetings have been instrumental in shaping this research.

I extend my sincere thanks to the committee members for their time and valuable feedback on this project.

I am immensely grateful to the members of the CPLS lab and the CEILing research group for their valuable feedback and collaborative brainstorming sessions. Their inputs have been helpful in shaping the direction of this research.

Special thanks to Professor Mary Bucholtz for organizing the Linguistics writing community. This community provided me with a focused environment and support to write my thesis effectively.

I would like to acknowledge and thank the Center for Scientific Computing (CSC) for providing the computing resources to run my programs.

I would also like to thank my partner, Anand Ramakrishnan, and my dog, Eva, for their unwavering emotional support during the entire duration of this project. Their presence and encouragement have been a source of strength.

I extend my thanks to my dear friends Lalitha, Maria , Kristy , and Cedar for their constant motivation and encouragement throughout this journey.

Last but not least, my deepest gratitude goes to my parents for their unending love, support, and belief in my abilities. Without their encouragement and support, I would not have reached this point in my journey.

ABSTRACT

Exploring the Extent of Statistical Learning used by Implicit Language Learners: Insights from Non-Māori Speakers Exposed to Māori

by

Ashvini Varatharaj

Recent works have demonstrated that New Zealanders who are frequently exposed to Māori in everyday life, but do not speak it, have an extensive memory store of Māori forms, called a proto-lexicon (Oh et al., 2020). This proto-lexicon is composed of morphs - words and word pieces that recur with statistical regularity in language usage that are learned through statistical learning (Ngon, et al., 2013). The proto-lexicon endows Non-Māori-Speaking New Zealanders (NMS) with rich implicit knowledge of Māori, which permits them to morphologically segment Māori words at above-chance levels (Panther et al., 2023a). Prior works (Saffran et al., 1996; Saffran 2003; Frank et al., 2013) have shown how statistical learning helps in implicit learning, but only in artificial languages. Oh et al. (2020) is one of the first studies to have shown this in real world exposure. In this work we use Morfessor (Smit et al., 2014), an unsupervised Bayesian segmentation model that identifies statistically recurrent morphs across words under the assumption of morphological concatenativity, to build on these recent studies to investigate the extent of statistical learning

used by NMS. We use Morfessor as our control statistical learner to perform two analyses. In our first analysis, we compare NMS and Morfessor to an expert Māori Speaker's (MS) ability to segment words into morphs. Comparing NMS and Morfessor's segmentation performances, we show the differences and similarities in the segmentation and learning process, and how it is affected by the statistical properties of the language. Further, using an error analysis on the segmentations, we gain insights into their underlying assumptions used in their segmentation process. The results of analysis 1 suggest that NMS may be sensitive to more than Morfessor, e.g. templates.

As a follow up to these results, in our second analysis, we dive deep into the results of the concatenative category of words whose structure closely resembles Morfessor's assumption. By generating pseudo-Māori words for this category and testing Morfessor's performance on them, we provide insights into how the statistical learning of real Māori morphs depends on explicit cues which it does not have access to – which the NMS seem to have some access to, where in they use the statistical regularities by taking a templatic approach in order to segment the words into morphs.

The most recent updated version of this work for publication can be found here : http://arxiv.org/abs/2403.14444.

TABLE OF CONTENTS

I. Introduction

Human beings have a powerful ability to learn language implicitly, through mere exposure and even without conscious effort or awareness. Implicit language learning refers to the process of learning a language by being exposed to it rather than a conscious effort to learn it. Prior research (Saffran, Aslin, and Newport 1996; Frank, Tenenbaum, and Gibson 2013) have shown how infants and adults are able to identify words in an artificial language after being exposed to it for a short duration through statistical learning, a process by which recurring structures in the language are identified and stored in the memory. Such learning has also recently been shown to occur outside of the lab, with real languages (Oh et al., 2020). The work by Oh et al. (2020) showed that Non-Māori Speakers (NMS) (– i.e New Zealanders who don't speak Māori but are frequently exposed to it in everyday life –), have implicit lexical and phonotactic knowledge. The phonotactic knowledge was best explained by the assumption that it derives from a memory store of phoneme sequences that recur with statistical regularity in the language, called *morphs*. Further, Panther et al. (2023b) replicated this result with more tightly-controlled stimuli and showed that an individual's lexical and phonotactic knowledge are correlated, lending crucial empirical support to Oh et al.'s assumption that phonotactic knowledge derives from lexical knowledge. Panther et al. (2023a) showed that NMS were able to segment words into morphs more accurately than Americans (who are not exposed to Māori) ie the segmentations made by NMS more closely mirrored those made by the fluent Māori speakers, supporting the claim that they have a memory-store of morphs. Furthermore, their segmentations were sensitive to phonological properties (such as phonotactics i.e. phoneme sequences which are more likely) in similar ways to those of fluent speakers, suggesting that the morphs learned by statistical learning align well with the actual underlying morphs of the language. These three works collectively provide evidence for the idea that non-native speakers (NMS) possess lexical and phonotactic knowledge, and that this knowledge is interconnected.

It has been postulated that this knowledge is gained through statistical learning, but relatively little is known about the precise aspects of linguistic structure that facilitate such learning. In this work we aim to take the first steps towards understanding this by focusing on two analyses. First, we compare the segmentations generated by Morfessor (Smit et al., 2014), a naive unsupervised morphological segmentation model on real Māori with the segmentations produced by Non-Māori speakers (NMS) exposed to real Māori. This comparison sheds light on the similarities and differences between the statistical learning processes employed by Morfessor and NMS. Second, we investigate the performance of Morfessor on actual Māori words compared to artificially generated Māori-like language (pseudo Māori). Using artificially generated languages provides us the capability to make sure it reflects the statistics of the language and that it follows Morfessor's concatenativity assumption i.e. that it does not contain other cues to morph boundaries that are in real Māori. By analyzing the differences between pseudo and real Māori, we gain insights into the limitations and deficiencies of Morfessor in capturing the complexities of real Māori. By putting together insights from these two analyses, we shed light on the similarities and differences between the statistical learning processes employed by Morfessor and NMS, contributing to a better understanding of implicit language learning through exposure. Furthermore, by examining Morfessor's performance across these distinct analyses, we aim to contribute to the broader implications of it towards unsupervised learning of morphological structure.

To summarize, the research questions of this work are as follows :

1. What are the similarities and differences between the statistical learning processes employed by Morfessor and NMS in the context of Māori language segmentation?
2. How does the performance of the Morfessor algorithm on actual Māori words compare to expectations derived from artificially generated Māori-like language (pseudo Māori), and what insights can this provide into the limitations and deficiencies of Morfessor in capturing the complexities of real Māori?
3. What are the implications for unsupervised learning of morphological structure in segmentation models, considering that Morfessor may perform differently in different subsets of data due to its potential limitations and missed assumptions? How can the understanding of these implications contribute to a better grasp of implicit language learning through exposure, particularly in the context of unsupervised morphological segmentation models when used in languages like Māori?

By addressing these research questions, the study aims to uncover the potential and boundaries of statistical learning in implicit language learning .

II. Background

A.Implicit language learning and Statistical learning

How humans learn to extract knowledge from their environment is one of the fundamental questions in cognitive science. Statistical learning refers to the process of extracting statistical regularities from input and adapting to them, based on considerations of frequency, variability, distribution, and co-occurrence (Saffran et al., 1996). Humans are highly sensitive to such statistical regularities and implicitly learn them from birth (Bulf, Johnson, and Valenza 2011; Gervain et al. 2008; Teinonen et al. 2009).

Implicit learning refers to the process of learning without intention, and even without the awareness of what has been learned (Williams, 2020) . Implicit learning plays a crucial role in human cognition as it underlies various essential skills such as language comprehension and production, intuitive decision making, and social interaction (Rebuschat, 2015).

A particularly prominent form of implicit learning is statistical learning. While early literature on statistical learning focused narrowly on transition probabilities, in this work "statistical learning" is being used more broadly to capture the learning of statistical morphological properties.

Saffran et al. (1996) was the first study that showed that infants were able to segment segment a fluent speech stream of a highly constrained artificial language into word-like units after just two minutes of exposure to the artificial language, thus showing how infants were performing word segmentation through statistical learning by being able to track the transitional probabilities. Estes et al. (2007) demonstrated that, for infants, exposure to word forms in a statistical word segmentation task facilitates subsequent word learning, thus showing the effect of implicit learning in the downstream task of word learning.

There have been other studies that have shown different regularities in language learnt through implicit statistical learning by children and adults. Saffran et al. (1997) investigated the word segmentation abilities of first-grade children and adults using an incidental language-learning task. Incidental-learning studies typically require that the subjects are engaged in a non-linguistic task while linguistic stimuli play in the background so that the participants are not actively and continuously attending to the linguistic stimuli. This way, any learning that occurred can be concluded as completely incidental, in that the attention was not directed to the language learning task which they are being tested for. The subjects were told that they were participating in an experiment investigating the influence of auditory stimuli on creativity. The experiment had adults and children perform a coloring activity on the computer (incidental task) while the auditory stimuli of artificial language with no pauses or any acoustic or prosodic cues to word boundaries was being played in the background. After the 20 minute experiment, the participants were given two sets of words and asked to guess which one sounded like the one that was played in the background during the experiment.The results showed that both age groups were able to learn the words of an artificial language presented in continuous speech, with children performing as well as adults.

B. Statistical learning in Real world languages

As seen above, previous research with artificial language learning paradigms has shown that infants are sensitive to statistical cues to word boundaries and that they can use these cues to extract word-like units (Saffran, 2001). However, this leads us to the question of whether infants perform statistical learning in real languages they hear outside the lab? Do they use this statistical information to construct word forms as they do in the artificially created languages? Pelucchi, Hay, and Saffran (2009) found that 8-month-old English learning infants, exposed to Italian, were capable of identifying patterns of transitional probabilities in language, even when the linguistic input was intricate, naturally spoken, followed grammatical rules, and conveyed meaning. Demonstration of statistical learning in a natural language allowed for greater ecological validity than previous experiments using artificial languages. However, these results tell us little about the representations that infants formed while listening to the fluent speech.

Ngon et al. (2013) showed that 11 month old French learning infants used statistical information to extract word candidates from their input. Using nonword stimuli matched with syllabic structure of real French words, the work showed that infants listen longer to high frequency disyllabic sequences than low frequency disyllabic sequences. This shows that infants are sensitive to the statistical property of frequency. Using another experiment comparing high-frequency nonwords and high-frequency French words as stimuli, they showed that infants showed no-difference between these two. Together this work depicted how infants, when they haven't yet learnt to segment words accurately from their input, are using the statistical recurrence of the units in their input towards the word-finding process. The units that recur sufficiently often are extracted as morphs and stored in memory, in a proto-lexicon (Johnson 2016). The proto-lexicon is a precursor to a fully-fledged mental lexicon: it contains forms, but not necessarily associated meanings, and may contain morphs corresponding to both words and non-words. From the above experiments, it is evident that infants, without conscious effort (implicitly) are extracting statistical properties in the input they are exposed to create their protolexicon. Thus we can say that the proto-lexicon acts as the seat upon which implicit knowledge of a language is built.

Another crucial limitation in the above words is that the exposure phase was performed in a lab based setup; while the stimuli were designed to be naturalistic, they are still highly constrained with respect to the number of words they are exposed to. Additionally, the infants were placed in sound attenuated booths and the setup was such that the infants would actively listen to the stimuli being presented. A crucial difference in the approach of the experiment involving non-native Māori speakers (NMS) as in (Oh et al., 2020), lies in the type of exposure they receive compared to this experiment. Unlike controlled experimental settings in a lab, their exposure occurs naturally in daily life situations, spanning a prolonged period. Additionally, the degree of engagement with the language varies among participants, reflecting a more organic and diverse range of interactions with the language, representing implicit learning scenarios in a real world setup.

Recent work by Oh et al. (2020) showed that Non-Māori speaking New Zealanders who have been exposed to Māori develop a Māori proto-lexicon through implicit statistical learning. Through a word identification experiment, the study showed that New Zealanders who are commonly surrounded by Māori but do not speak it were able to distinguish real Māori words from word-like (phonotactically-matched) nonwords, thus demonstrating implicit lexical knowledge. Further, using a wellformedness rating experiment, participants rated Māori-like nonwords for how good they would be as a real Māori word, using a scale ranging from 1 ('Non Māori-like non-word') to 5 ('Highly Māori-like non-word'). The words for the stimuli represent different degrees of phonotactic wellformedness based on the statistics of the language. The participants involved three groups - NMS, fluent Māori speakers, and non-Māori speaking Americans. The Fluent Māori speakers provide us with a baseline metric given that they have full lexicon and a complete phonotactic knowledge. The US participants are at the other end of the comparison, given that they have no lexicon and almost no phonotactic knowledge of the language. The results showed that the US participants showed only a slight increase in their ratings across varying levels of phonotactic probability, suggesting they have minimal specific knowledge of Māori phonotactics. In contrast, both NMS demonstrated a substantial increase in their ratings in line with phonotactic probability, indicating a significant understanding of Māori phonotactics. This shows that the ratings provided by NMS is not by mere guessing, rather it is a result of their exposure to Māori, indicating proof of non-negligible phonotactic knowledge of Māori. Additionally, the performance of the NMS did not significantly differ from that of MS, which shows that NMS have gained a strong phonotactic knowledge of Māori which is similar to that of MS.

To analyze the source of NMS' phonotactic knowledge, simulations of proto-lexicon with varying vocabulary sizes of Māori lexicon were used to predict NMS' well-formedness ratings. The results showed that the well-formedness ratings of NMS participants can be adequately explained by assuming that their phonotactic knowledge is based on a proto-lexicon consisting of 3,000 common Māori words. Further, the authors noted that from what is known about statistical learning, it is possible that the NMS proto-lexicon does not consist of words at all, but is rather made up of morphs, which are phonological (sub)sequences that recur across different words. To test the cognitive assumption that the NMS proto-lexicon consists of morphs, phonotactic probabilities were calculated based on morphs (obtained from morph segmentation of words by a fluent Māori speaker). Using ordinal mixed-effects regression models, the morph-based phonotactic probabilities better predicted NMS's ratings than the word-based phonotactic probabilities, thus suggesting that the NMS' protolexicon most likely consists of morphs. Using simulations of a proto-lexicon consisting of morphs, the work showed that the NMS participants' ratings can be adequately predicted by phonotactic knowledge generated over a set of approximately 1,500 of the most common morphs.

Through modeling of these results, it was concluded that the best fit of the performance of these non-Māori speakers in the lexical tasks is explained by their statistical learning of word parts or morphs, which involves segmenting words into smaller components and storing them for future use. Building on this, recent work Panther et al. (2023a) showed that morphological segmentations by Non-Māori speakers in New Zealand matches the segmentations by proficient Māori speakers, thus adding to the literature that NMS are able to gain speaker-like knowledge through ambient exposure and implicit statistical learning.

It brings up the important and interesting question of how language structure affects the learning by non-speakers of a language to segment words in a real language and process morphological complexities of the language. Todd et al. (2023) raise this question in a replication of Oh et al.'s (2020) work, targeting implicit knowledge of Spanish held by non-Spanish-speaking Californians and Texans. In this work, the authors showed that non-Spanish speakers in California and Texas (states where Spanish is largely spoken), have implicit lexical and phonotactic knowledge of Spanish. However, it appears to be weaker than the knowledge of Māori held by Non-Māori speakers in New Zealand studied by Oh et al. (2020). One potential explanation is the structure of the language , morphology being a notable structural difference – Spanish has morphological differences to Māori, such a lower use of compounding compared to (inflectional or derivational) affixation.

C. Morphological Segmentation

Having seen that adults use phonotactic cues to segment words from sentences, and knowing that language contains recurring structures, it is highly likely that learners are implicitly learning these structures in the process (Panther et al., 2023a). The literature on modeling morphological segmentation processes reinforces this perspective, indicating that morphological segmentation can be accomplished without relying on semantic knowledge. Various algorithms have been proposed, demonstrating that recurrent morphological patterns can be statistically learned in a bottom-up manner solely from exposure to word forms (Creutz and Lagus, 2007; Daland and Pierrehumbert, 2011). Some algorithms like Morfessor (Creutz and Lagus, 2007) assume that segmentations require reference to an inventory of morphs, so there is an assumed proto-lexicon. Whereas other algorithms such as DiBS (Daland and Pierrehumbert, 2011) assume that segmentation can occur based on phonological transition probabilities alone, without the need of a known inventory (proto-lexicon).

In morphological segmentation the goal is to identify boundaries within words by splitting them into morphemes, the smallest meaning-carrying units. In unsupervised approaches, the inventory of parts is inferred from the training data, by identifying the morphs – sequences of characters, phonemes, or larger ‘atoms’ – that recur across words with statistical regularity. For this work we use Morfessor (Virpioja et al., 2013)) as our naive statistical learner, a generative probabilistic unsupervised morphological segmentation model. We chose to use Morfessor due to its simple assumptions and is often used as a baseline among morphological segmentation models.

Unsupervised morphological segmentation provides us an avenue to simulate implicit statistical learning processes. In this work , we use Morfessor , which is one of the popular unsupervised morphological segmentation models in the field. Morfessor operates within a Minimum Description Length framework (Rissanen, 1978), aiming to identify the most concise and straightforward set of morphs (the lexicon) that can generate the training data with the highest probability. In this approach, the lexicon is treated as a collection of morphs, and during training, the cost of adding a particular morph to the lexicon is determined based on both its complexity and how frequently it appears across words. The training data are assumed to be generated from the lexicon by concatenating morphs drawn independently from it, without considering any constraints related to their position, sequencing, or morphosyntactic category. Morfessor is based on the assumptions that words are composed of morphs, that frequent morpheme sequences indicate valid morphological units, and that language users are capable of generating new words through productive morphological processes. This then neatly relates to the assumption which was modeled in Oh et al. (2020), which showed that NMS's phonotactic knowledge can be best explained by a protolexicon made up of morphs. Morfessor thus can be a good candidate to understand the exact underlying implicit learning processes used by Non-Māori speakers (NMS). By drawing parallels between the learning processes between NMS and Morfessor, this work aims to look into the extent of statistical learning used by NMS.

In the context of our discussion, it is essential to acknowledge the inherent limitations of Morfessor, which arise from the simplifications entailed by its assumptions. However, as we will elaborate in the following section, the structural characteristics of the Māori language align well with these assumptions. This alignment suggests that Morfessor can be considered a plausible naive cognitive model for the purposes of our analysis.

III. Māori Language

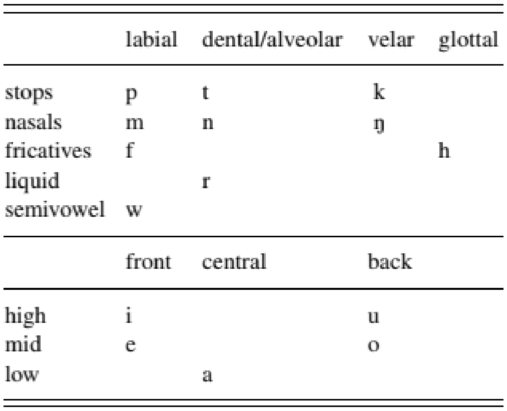
Māori, or te reo Māori, commonly shortened to te reo, is an Eastern Polynesian language spoken by the Māori people, the indigenous population of mainland New Zealand. As of the latest data available, the 2018 New Zealand Census reported that there are 185,955 Māori speakers which accounts for approximately 3.9% of the total population of New Zealand.

The Māori phoneme inventory consists of five vowel and ten consonant phonemes as show in Table 1. The ten consonant phonemes are : /p, t, k, m, n, ŋ, w, f, r, h/ and the five vowels are /i, e, a, o, u/. The orthography in Māori is highly transparent i.e. the written form closely resembles the spoken form. The consonants are represented by <p, t, k, m, n, ng, w, wh, r, h>, and the vowels represented by <i, e, a, o, u> respectively. Vowel length is phonemic, and each vowel has a long counterpart. Long vowels are represented with a macron: <ī, ē, ā, ō, ū>. Māori syllables follow a (C)V(V) template, with optional simplex onsets and no codas. Māori also has a transparent morphological system, which consists of little inflectional and derivational morphology, and in which compounding is frequent (Harlow, 2007). This makes it suitable for morphological segmentation models such as Morfessor.

In the context of the Māori language, there is not (yet) a consensus on what is and is not a diphthong, based on phonetic properties (e.g., the absence of hiatus as you mention) or phonological properties (e.g., the influence on stress assignment). Furthermore, the status of a VV sequence as a diphthong or sequence of monophthongs is affected by morphological structure, and it is not yet clear the extent to which diphthongization across morpheme boundaries occurs. Given these outstanding questions, we follow other quantitative work on Māori in treating all Vs separately.

Thus, for the purpose of this study, we only model CV and V structures. The models hence are not particularly distinguishing diphthongs and sequences of monophthong structures as it could lead to ambiguity. The table below (Harlow, 2007) shows the phoneme tables.

Table 1. Phonemes in Māori



	labial	dental/alveolar	velar	glottal
stops	p	t	k	
nasals	m	n	ŋ	
fricatives	f			h
liquid		r		
semivowel	w			
	front	central	back	
high	i		u	
mid	e		o	
low		a		

IV. Analysis 1: Comparing Morfessor and Non-Māori Speaker (NMS) segmentations on real Māori words.

Non-Māori speakers in New Zealand exposed to Māori are similar to Morfessor in the sense that both are learning to segment the language based on statistical patterns in the language they are exposed to without getting feedback (unsupervised). In both cases the learners are using recurring units to learn word parts in order to build their proto-lexicon. By comparing them, we can understand how similar or different these two learning processes are which will help understand how statistical learning plays a role in implicit language learning through exposure.

A. Data

The dataset used here is the NMS **Non-Māori Speaker (NMS) data** and the **MS (Māori Speaker) data**.

1. Non-Māori Speaker (NMS) data: Non-speaker segmented data is used from Panther et al. (2023a) where 195 non-Māori speakers in New Zealand were asked to segment Māori words into morphs. During the experiment, the participants were given multiple examples of morphological complexity in English and Māori before providing the task of segmenting Māori words. The stimuli were presented orthographically. They were asked to click between any two letters to assign a segmentation or to click a box under the word to leave the word unsegmented. Each label obtained from the participants indicates whether or not the participant thought if there was a boundary between the two letters.

Aggregating NMS data : In order to use the NMS data as a comparison between Māori Speakers as well as Morfessor, it needed to be aggregated. For each letter pair in each word, there were labels from multiple participants where the labels indicated if each participant thought whether there was a boundary between the two letters. To aggregate at the word level, the majority of the labels at each position were taken as the final label for each word. For example, Table 2. below shows the aggregated data for one word ‘hoiho’.

Table 2. Example of aggregation of labels from NMS data.

word	true_votes	false_votes	majority	segmentation
hoiho	[0, 1, 8, 0]	[11, 10, 3, 11]	[False, False, True, False]	hoi+ho

The participant labels were counted based on whether they placed a boundary between each pair of letters. For the word ‘hoiho’, there are 4 possible places where they could place a boundary – between the letters ‘h’ and ‘o’, ‘o’ and ‘i’, ‘i’ and ‘h’, ‘h’ and ‘o’. The true votes are the counts of how many participants placed a boundary in these positions in the given order. Similarly false votes are the number of participants who did not place a boundary in that position. The majority is taken between the two sets of votes to determine the aggregated segmentation which is ‘hoi+ho’. There are a total of 4427 words in this dataset.

2. Māori Speaker (MS) data :We use the word segmentation data collected from a fluent Māori expert speaker (MS data) in Oh et al. (2020). As described in their work, the initial corpus consisted of 19,595 words from the Te Aka dictionary (Moorfield, 2011). However, since segmentations could reliably and straightforwardly be inferred for some words – either because they are too short to be complex or the result of productive and transparent morphological processes, these words were excluded from the words given to the speaker. 1,014 words that were identified as simplex—bimoraic or smaller—were excluded. Additionally, 34 bimoraic words composed of a repeated syllable were selectively evaluated for potential segmentation. Another 6,360 words were held out in order to reduce redundancies, so that the raters didn't have to rate more than necessary in order to confidently arrive at segmentations for all words in the dictionary. The rationale behind this selection was that these words were formed by using a transparent morphological process to a base stem, which speakers were already breaking down into segments. And once the segmentation of the stem were known, the segmentation of the held-out word could be inferred from it. The remaining 12,221 words were presented to the fluent speaker for segmentation.

For the 7,374 words initially held out from decomposition, inferences were made based on the decompositions of related words within the dictionary. Specifically, the 1,014 short words presumed to be simplex were inferred to contain only a single morph. For the 6,360 words that were likely products of transparent morphological processes using known stems, their decompositions were inferred by applying the known morphological rules to the identified stems, adding affixes and reduplication where appropriate. Words formed from stems that the speaker could not recognize were marked as unknown, which accounted for 71 words. As a result, segmentations for a total of 19,524 words were obtained. Further methodological details can be found in the supplemental document of Oh et al. (2020).

When comparing MS and NMS performances, we take a subset of the data to match the words from the NMS data to calculate the performance metrics.

B. Method

Morfessor segmentations : Morfessor was trained and tested on the words for which we have the NMS segmentations i.e the 4427 words. We obtain the segmentations for those words for which we have the word category information in order to compare the different word categories across the two learners.

The broad categories of words in the dataset are **polymoraic**, **affixation**, **monomorphemic** and **reduplication** .

1. The **polymoraic** words category encompasses terms that consist of four or more moras; these words may not display transparent morphology yet could possess complex structures. 2. **Monomorphemic** words consist of a single morpheme.

   a. Bimoraic disyllable words are a subset of monomorphemic words that are composed of two syllables, each typically containing a short vowel.

   b. Trimoraic words, another subset, consist of three moras and may feature different syllabic structures, including long vowels.

3. **Affixation** category words are those that carry one or more affixes and are further subdivided into three specific types:

   a. Nominal,

   b. Passive, and

   c. "whaka" prefixed words that typically denote a causative action in the language.

4. The **reduplication** category involves words that exhibit repetition of word parts, either partially or in full. This category itself has several subcategories:

   a. partial reduplication_left short where the first mora (syllable containing a short vowel) is repeated to the left,

   b. left-reduplication where 2 morae in a word with more than 2 morae are repeated to the left,

   c. total reduplication where the entire word is duplicated to create a new meaning.

   d. partial reduplication_left long which involves the first syllable of the base repeated to the left (and has its vowel lengthened) , and

e. partial reduplication_right where a portion of the end of the word is repeated.

The sub categories of words and their counts are shown in table 2 below.

Metrics : The metrics used to compare performance are boundary precision and recall (Stolcke and Shriberg, 1996). In our use case of identifying morphological boundaries in words, precision and recall can be understood in the context of whether the boundaries are identified correctly by the two learners (NMS and Morfessor) compared to the true morphological boundaries. It is to be noted that the "true" boundaries are taken to be those produced by the fluent speaker and we acknowledge that there might be variation between speakers for these boundaries. Precision in this context refers to the proportion of the boundaries identified by the learner that are actually correct. Recall refers to the proportion of the true boundaries correctly identified by the learner. For example, if we have a word that should be segmented as A+B+C and the provided boundary position by NMS is A+BC, then the precision in this case would be 1 or 100% and the recall would be 0.5 or 50% since the learner missed one boundary. The precision and recall is calculated at the word level and averaged for each subcategory.

Edge cases: In the context of this analysis, we have adopted specific conventions to handle edge cases where traditional precision and recall metrics may not be directly applicable:

1. **True Zero Boundaries:** If the correct segmentation contains no boundaries, and the learner also predicts no boundaries, we define both precision and recall as 1. This reflects perfect agreement between the learner's predictions and the ground truth, even in the absence of boundaries to detect.

2. **Missing Learner Boundaries:** When the true segmentation contains boundaries, but the learner fails to predict any, traditional precision would be undefined due to a zero divisor. For the purposes of analysis, we will consider precision to be 0 in this scenario. This decision aligns with the principle that the learner has failed to detect any of the true boundaries it was supposed to find.
3. **Incorrect Learner Boundaries:** In the case where the true segmentation has no boundaries (e.g., mono-morphemic words) but the learner incorrectly predicts boundaries, recall would traditionally be undefined. Again, for the sake of consistency in analysis, we will treat recall here as 0. This reflects the learner's error in predicting boundaries where there should be none.

While these conventions might affect the interpretation of the results, they are necessary to ensure that the analysis remains coherent and can accommodate all possible scenarios. We want to note that we are aware of the limitations inherent in this approach during interpretation of overall results. By following up with an error analysis alongside our quantitative measures, we aim to address these limitations and provide a more comprehensive understanding of the learner's performance in this context.

Table 3. Precision, recall for each of the word categories

category	morf_prec	NMS_prec	morf_rec	NMS_rec
polymoraic	0.70	**0.8**1	**0.88**	**0.82**
monomorphemic	0.26	0.71	0.28	0.72
affixation	**0.72**	0.72	0.86	0.72

reduplication	0.61	0.77	0.83	0.79

C. Results

Table 3. shows the comparison of average precision, recall for Morfessor and NMS for each of the four categories of words. Overall the NMS speakers seem to be performing better than Morfessor in all categories in terms of precision with the affixation category performance being equal to Morfessor. NMS precision is the highest in the polymoraic category followed by reduplication, affixation and monomorphemic categories. However, with Morfessor, the highest precision is obtained in the affixation category followed by polymoraic, reduplication and monomorphemic categories. The recall metric however has a different trend from the precision. Morfessor seems to consistently have a higher recall in all the categories, the highest being in polymoraic category. This could be due to over segmentation by Morfessor. NMS has the highest recall in the polymoraic category as well. NMS recall is the lowest in the monomorphemic and affixation categories. We will dive deep into the results of each of these categories and subcategories next. Table 4 below shows the sub-category of words and their respective precision,recall metrics.

Table 4. Precision, recall and average frequency of the morphs each learner is exposed to for each category.

category	**sub category**	**morf_prec**	**morf_rec**	**NMS_prec**	**NMS_rec**

monomorphemic	bimoraic_disyllable	0.21	0.21	0.88	0.88
	trimoraic	0.31	0.35	0.54	0.55
affixation	nominal	0.71	0.84	0.66	0.67
	passive	0.67	0.77	0.56	0.56
	whaka	0.77	0.97	0.95	0.93
reduplication	total_redup	0.55	0.88	0.95	0.97
	partial_redup_left_long	0.69	0.95	0.8	0.82
	partial_redup_left_short	0.45	0.6	0.55	0.59
	partial_redup_right	0.76	0.88	0.78	0.77
polymoraic	polymoraic	0.7	0.88	0.81	0.82

Monomorphemic words : Any word with three or fewer morae is monomorphemic. Monomorphemic words as mentioned can be either bimoraic_disyllable or trimoraic. For example ‘pewa’ is a bimoraic_disyllable. NMS predicted ‘pewa’ with no boundaries, while Morfessor predicted ‘pe+wa’. While both the learners seem to be struggling, Morfessor is better in the trimoraic category than the bimoraic disyllable, whereas NMS appears to be performing substantially better than Morfessor in both the categories, bimoraic di_syllable being its highest in both precision and recall.

Affixation : Within the affixation category, both Morfessor and NMS struggle the most with nominal subcategory and perform the best in whaka, although the Morfessor outperforms NMS in the nominal and passive categories. This can be attributed to the fact that “whaka” contains the most common prefix, and so it should be easily extracted by both Morfessor and human learners. To understand the performances in nominal and passive categories, we need a fine grained analysis of the default and non-default sub categories within these affix categories. Table 7 below shows the split of default and non-default affixes in the nominal and passive categories. The nominal category has ‘haNa’ and ‘taNa’ as the default allomorphs of the -Canga affix. The non-default variants of it are 'aNa', 'Na', 'kaNa', 'maNa', 'raNa', 'faNa'. Similarly the passive category has ‘tia’ and ‘hia’ in the default allomorphs of the -Cia affix category and has variants of the non-default category: ‘Nia', 'a', 'ia', 'ina', 'kia', 'mia', 'na','Na', 'ria', 'fia', 'fina', 'kina'.

To compare how the statistical recurrence of default and non-default affixes in the input affects the performance of these learners in each of the word categories, we create the average morph frequency and affix accuracy metric.

Average frequency : The average morph frequency metric is designed to quantify the exposure frequency of various morphs to learners. For Morfessor, the average morph frequency calculation is based on the morphs present in the word types that are input into the model, with a reference to the segmentation provided by an expert Māori Speaker (MS). In the case of NMS, the average morph frequency metric is derived from their exposure to the morphs present in the word types in the Te Aka dictionary (Moorfield, 2011). It was calculated by dividing the total number of word types each morph occurred in by the sum of all such counts for all the morphs. This acts as a proxy for the linguistic input available to NMS.

The average morph frequency at the word level is calculated by taking the average of the corresponding morph frequencies contained in the word. This process essentially provides a measure of how frequently Morfessor and NMS encounter specific morphs in the language across different word categories. By comparing these frequencies, we can gain insights into the differences in morph frequencies in relation to their morphological segmentation process. Removing the stem and looking at the affix frequency alone can give us a better idea to interpret the results of the different sub categories. To understand the effect of affix segmentation alone, we created an affix_acc metric to calculate the accuracy of segmentation of the affix in each category as shown in the table 5 below. The affix_acc ranges from 0 to 1. The Figure 1 below shows the distribution of default and non-default allomorphs for the different affix categories.

Average affix accuracy : This metric is based on the accuracy of getting the affix segmentation correct. Irrespective of the remaining morphs in the words, the affix accuracy for a word is 1 if the affix has been correctly segmented. The average of the affix accuracies for all the words in the category is presented as the value for each category in Table 5.

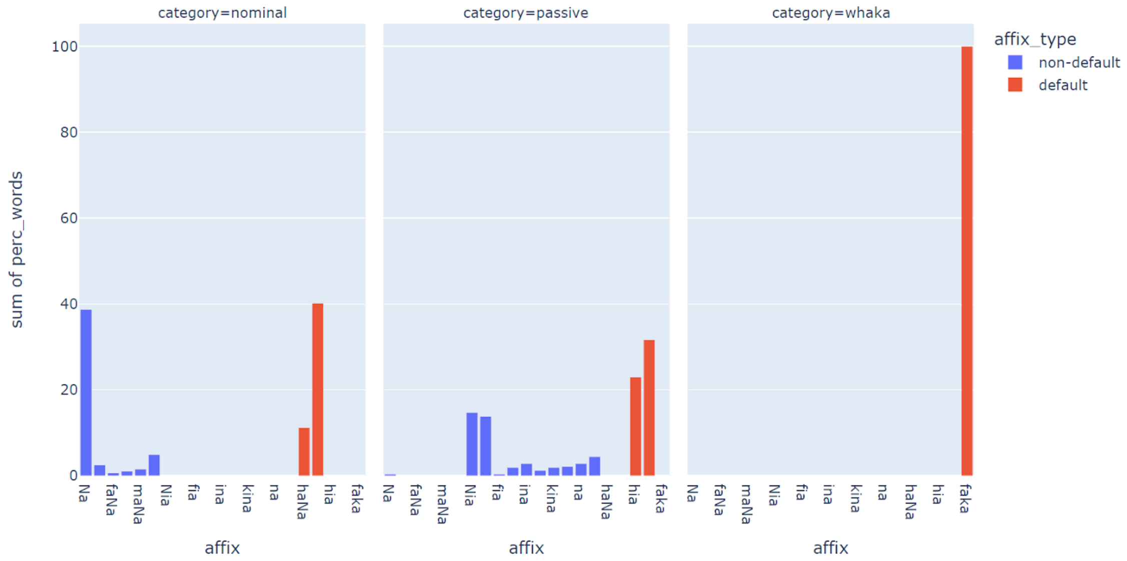


Figure 1. Distribution of non-default and default affix categories among nominal, passive and whaka affix categories.

Table 5. Affix segmentation accuracy and frequencies for Morfessor and NMS

			Morfessor		NMS	
category	affix_type	affixes	affix_acc	affix_frequency	affix_acc	affix_frequency
nominal	default	haNa','taNa'	1.00	0.97	0.89	0.77
	non-default	aNa', 'Na', 'kaNa', 'maNa', 'raNa', 'faNa'	0.88	1.15	0.53	0.66
passive	default	tia','hia'	1.00	1.58	0.72	3.86
	non-default	Nia', 'a', 'ia', 'ina', 'kia', 'mia', 'na','Na', 'ria', 'fia', 'fina', 'kina'	0.77	0.63	0.34	0.79
whaka	default	whaka	1.00	3.61	0.98	6.88

Nominal Category:

- *Default Affixes*: In nominals, it looks like within the default affix category both Morfessor and NMS are doing relatively well. The high affix frequency of 1 in Morfessor and 0.89 in NMS attests to this. This could be due to the fewer items of affixes in this category along with the fact that the affix frequency is high for this category.

- *Non-Default Affixes*: Here ,the performance drops for both Morfessor and NMS, with Morfessor maintaining a lead. Although the mean affix frequency for this category is higher, the greater variation could be the reason behind the drop in affix accuracy in Morfessor. In the case of NMS, the affix frequency drops which, along with the multiple variations could cause the lower affix accuracy. This suggests that the lower frequency and greater variation in non-default affixes pose challenges to effective segmentation in both learners. In the case of Morfessor, the recall is high, which can be attributed to the over segmentation behavior. One possible explanation for the low performance by NMS can be that the affix frequencies haven't reached a threshold for them to confidently identify the different templates needed.

Passive Category:

- *Default Affixes*: Similar to the nominal category, default affixes see better performance compared to non-default categories. Morfessor has a high affix frequency within this category which could be the reason underlying the high accuracy in segmenting these default affixes. NMS on the other hand, has a relatively high affix frequency although this is not leading to the same increase in its affix accuracy. Further fine grained analysis of the individual affix performance might shed light on this behavior.

- *Non-Default Affixes*: Both Morfessor and NMs have the lowest affix accuracy in this category. NMS appears to be struggling more than Morfessor. The complexity of the multiple allomorphs could be an important contributing factor. Similar to the non-default category in the nominal affixes, the low performance by NMS can be that the affix frequencies haven't reached a threshold for them to confidently identify the different templates needed.

'Whaka' Category: The whaka category results seem to be straightforward. Both NMS and Morfessor have a high frequency of whaka in their input which could result in the correct segmentation of the affix in almost all the words (accuracy of 1 and 0.98 in Morfessor and NMS respectively).

While there are some similar trends across the subcategories, NMS and Morfessor seem to be learning through different processes by using the statistical regularities in distinct ways which needs further analysis by breaking down each of the non-default affix performances as well.

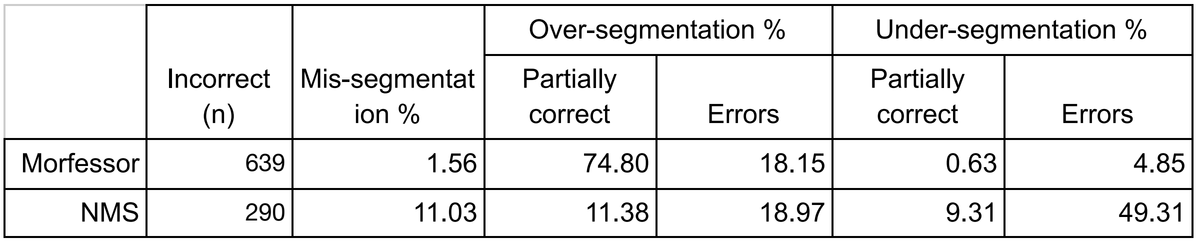
Reduplication : Morfessor is not able to learn patterns like humans. NMS speakers are able to identify reduplicated patterns and so have a much higher precision compared to Morfessor. Within reduplication subcategories, it is interesting to see how the total_redup frequencies are the lowest for Morfessor and NMS within this subcategory, but NMS seem to be having the highest precision in total_redup, indicating an intuitive process which seems to be helping NMS segment these words. However, for Morfessor, which depends on the distributional properties of morphs, doesn't seem to be picking up on recurrent patterns.

We can see that partial_redup_left_short (Leftward Redup with short vowel) is the hardest for both Morfessor and NMS in terms of precision. Morfessor seems to be performing better in the partial_redup_right than in the partial_redup_left_long, even though the average frequency is higher in the later category. Further analysis on the different patterns within this category might help shed light on this behavior. NMS on the other hand , have a better precision for partial_redup_left_long than partial_redup_right.

Polymoraic: These are words which contain four or more moras and are the class of words without transparent morphology. These are the categories of words that don't contain reduplication nor affixations. Overall, It appears that NMS are performing relatively better than Morfessor in terms of precision(prec_Morfessor = 0.70, prec_NMS = 0.81) whereas Morfessor's recall is better than NMS (prec_Morfessor = 0.82, prec_NMS = 0.88). While it has a lower precision, it has a high recall which could be due to Morfessor oversimplifying the segmentations by either over-segmenting or mis-segmenting. The error analysis in the next section shows the distribution of over-segmentation, mis-segmentation and under-segmentation by Morfessor and NMS for the Polymoraic category.

Polymoraic words error analysis : Figure 2 shows the errors for the polymoraic word category.

Table 6. Error analysis on polymoraic words



	Incorrect (n)	Mis-segmentation %	Over-segmentation %		Under-segmentation %	
			Partially correct	Errors	Partially correct	Errors
Morfessor	639	1.56	74.80	18.15	0.63	4.85
NMS	290	11.03	11.38	18.97	9.31	49.31

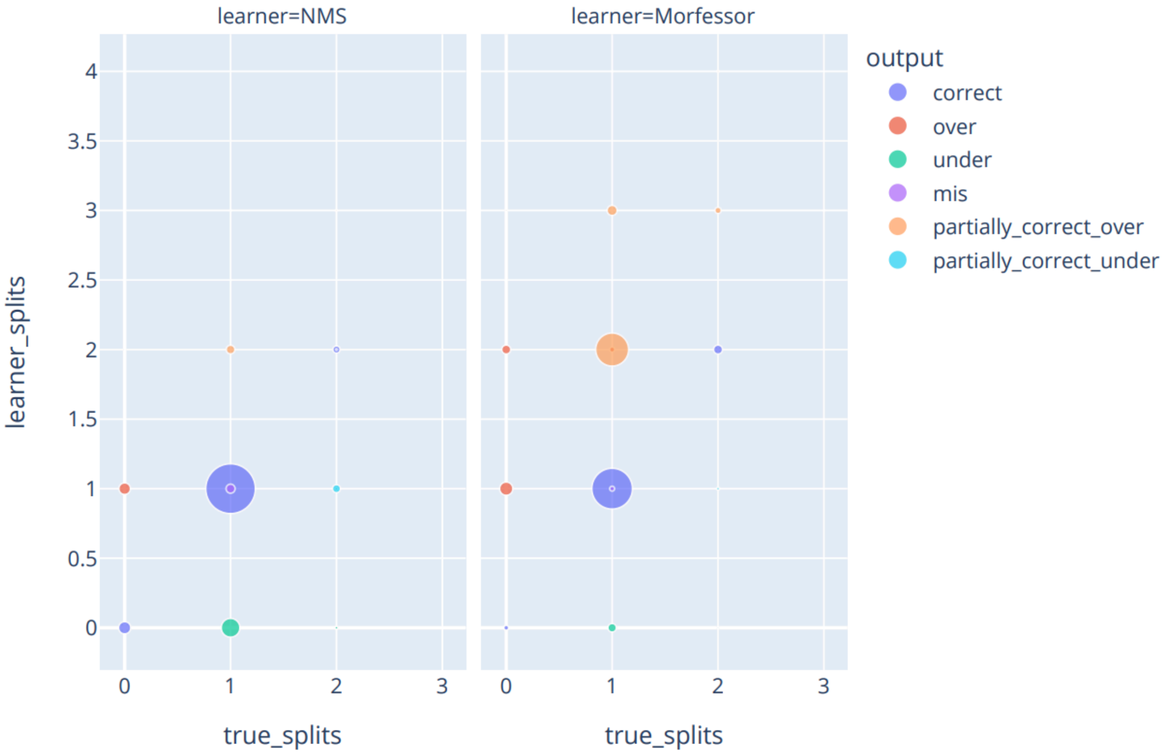


Figure 2. The distribution of correct splits vs the splits made by the learners categorized as correct segmentations, over, under, partially correct and mis segmentations in polymoraic words.

The figure shows a visual representation of the different kinds of errors made by the learners while the table shows the quantitative analysis of the error types. We analyze the errors drawing from these two representations. We define partially correct over-segmentations to see if at least part of the oversegmentations contain actual correct morphs ie even though the boundaries are not placed in the correct positions, we check if there is a subset of correct morphs with the provided boundaries Similarly, we check if there is a subset of correct morphs in the over segmentation cases, which we define as partially correct over segmentations. This gives us an idea that while the entire word was not correctly segmented in all the positions, a part of the segmentations were indeed correct.

Table 6. shows the error analysis on the polymoraic words. Morfessor, tends to have a significant number of over-segmentations, accounting for approximately 92.96% of the errors. However, a large majority of it (74.80%) consists of partially correct segmentations i.e out of the oversegmentations, some of the boundaries are being placed in the correct positions. Over-segmentation occurs when Morfessor divides words into smaller units or morphemes more frequently than necessary. This high rate of over-segmentation contributes to the high recall, as Morfessor captures a large portion of the true morphs in the words. However, it also introduces incorrect morphs thus affecting the precision as seen above.

The next category of errors made by Morfessor is under-segmentation, which accounts for around 5.48% of the errors. Out of these, a small percentage do contain partially correct morphs (0.63%). Under-segmentation occurs when Morfessor fails to identify an adequate number of morphemes within a word. This can happen due to various reasons, such as infrequent occurrence of individual morphemes or the limitations in Morfessor's assumptions about the structure of the language or corpus which we analyze to some extent using the next analysis.

Mis-segmentations are the least common type of error made by Morfessor, representing approximately 1.56% of the errors. Mis-segmentation refers to cases where Morfessor incorrectly splits a word into morphs, resulting in incorrect boundaries between them.

On the other hand, the most common error made by NMS is under-segmentation, accounting for around 58.62% of the errors with 9.31% of them containing partially correct morphs. This suggests that the NMS have difficulty identifying an adequate number of morphs in the words, possibly due to having a high threshold of confidence required before they commit to a boundary being present. The second most common error made by NMS is over-segmentation, representing approximately 30.34% of the errors with 11.38% being partially correct. It is worth noting that both Morfessor and NMS exhibit fewer errors in the mis-segmentation category. NMS have a higher error rate of 11.03% compared to the 1.56% by Morfessor.

To summarize, Morfessor tends to have a higher rate of over-segmentation errors, followed by under-segmentation and mis-segmentation errors. In contrast, the NMS show a higher rate of under-segmentation errors, followed by over-segmentation and mis-segmentation errors. These findings provide some initial insights into the strengths of the processes each takes in identifying morphs and where they lack.

Findings from the above results exemplify how two learners that are similar in their learning processes show significant differences between them in their performances. Results from Analysis 1 indicate that overall NMS are better than Morfessor in the morphological segmentation task. It seems like there are cues to morphological segmentations in Māori which NMS are using, that go beyond simple recurrence statistics as utilized by Morfessor, e.g knowledge of affixation and reduplication templates. If this is the case, then we expect that Morfessor would do much better on a language that followed exactly the same morphological statistics as Māori but lacked sensitivity to these other cues. To verify this assumption, we create artificial languages resembling Māori but adhering strictly to Morfessor's assumed language structure. If Morfessor is able to perform well on pseudo-Māori, then it suggests that the learning mechanisms NMS use are far more complex than the one used by Morfessor.

V. Analysis 2 : Morfessor performance on pseudo-Māori polymoraic words

From Analysis 1, it appears that NMS are utilizing cues for morphological segmentation in the Māori language that extend beyond the basic recurrence statistics employed by Morfessor. These cues might include factors like specific patterns such as the reduplication templates. Given this observation, it's reasonable to hypothesize that Morfessor would perform more effectively with a language that has morphological statistics identical to Māori but does not require sensitivity to these additional cues. We test this hypothesis by generating pseudo-Māori words which are highly constrained to follow the morphological statistical properties of real Māori words.

From the different word categories, we create pseudo-Māori words only for the polymoraic category in this analysis since the polymoraic category most neatly fits the generative process that Morfessor assumes. Each other category has some limitation that makes it ill-suited to this analysis. For instance, the reduplication category of words, as seen from Todd et al. (2022), would benefit from Morfessor having reduplication templates (which we are not using in the current analysis). The Affixation category would need some way to inform Morfessor about the allomorphy i.e the fact that there are multiple variants of the same affix. Monomorphemic words, by definition, have no boundaries in them; so the analysis would be limited to identifying error cases where Morfessor predicts a boundary when there isn't one.

A. Data

Pseudo-Māori Generation : The pseudo Māori language generator was a simple generative model built in python 3.9.7. This involved a two-step process. First, we obtained the parameters from real Māori words in order to simulate those properties in pseudo Māori. The next step involved generating the words in the pseudo Māori language using the parameters obtained in step one. Each of these steps are explained in detail below.

We use a generative process to generate the pseudo-Māori words which closely resemble the real Māori words. To develop words which have properties similar to the real Māori words, we need to obtain the statistical properties of the word at different levels (syllable, morph, word) in order to create the pseudo words. So we first calculate these properties in the form of parameters and then use them in our generative process.

Below are the steps involved in generating the pseudo-Māori words for the polymoraic category.

B.Method

1. **Getting parameters from real Māori words (polymoraic category):** This process involved two main steps: first, identifying the statistical parameters from the words at the morph level in the Māori language for the polymoraic words. The second step involved calculating parameters at the syllable level. To extract parameters, we use the polymoraic words from the MS data for which we have the word category information. This consists of 1317 words.

   A. **Parameters at the morph level:** For each word in this subset Māori data, we create a pseudo-Māori word with similar statistical morph properties. First, we calculated the number of syllables present in each morph of a word. For example if a word contains two morphs, and the number of syllables present in each morph is two and three respectively, then we store that as the syllable count ([2,3]). To simplify, we filter out any words containing morphs made up of four or more syllables since there were only 22 words in the MS dataset fitting that criteria. This decision is also grounded in the desire to maintain a manageable complexity within the model while capturing a representative range of morphological variability. We calculate the type frequency of each morph i.e the number of words it occurs in. We model the distribution of type frequencies across morphs separately for mono-syllabic, disyllabic and trisyllabic morphs. We then group mono-syllabic morphs, di-syllabic morphs and tri-syllabic morphs to fit a power law on each of these three distributions since Morfessor assumes a power law distribution.

The distribution of each morph category was modeled using a power law and the parameters of the power law are estimated using curve fitting(with Scipy's curve_fit) (Virtanen et al., 2020). Scipy's curve_fit uses nonlinear least squares to fit a function, f, to data. It returns optimal values for the parameters so that the sum of the squared residuals of the function f is minimized. In this case the function f is a power law whose equation is:

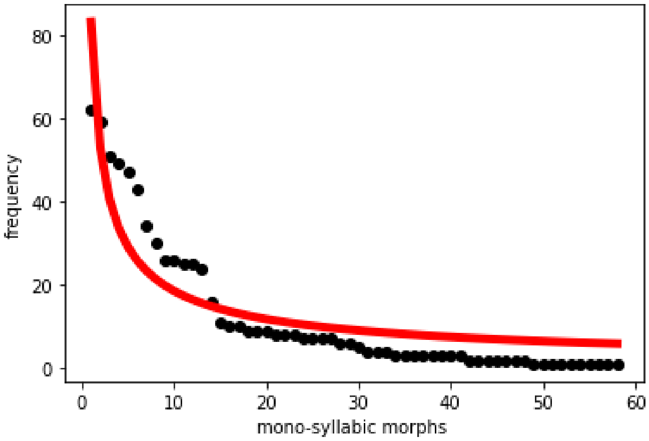
$$y = a * b^{x}$$

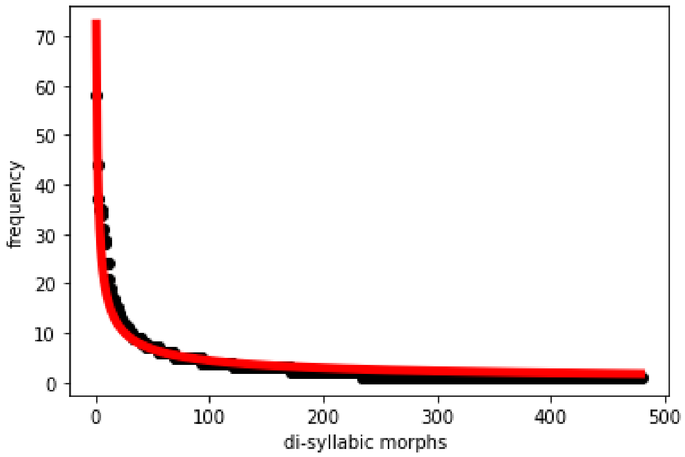
The curve_fit algorithm returns two arrays: **params** - An array with the optimal values of parameters **'a'** and **'b'** that it found for the power function. Figure 3 below shows the fitting of the power law for the three categories of morph distribution.

We obtain three sets of such parameters, one for each category, which we then use to generate pseudo mono,di, tri-syllabic morphs. The number of morph types in these categories are based on the counts present in the subset data. Table 7 below shows the morph type counts and the parameters obtained for each of the three categories of morphs.

Table 7. Morph type count and power law parameters for mono,di,tri-syllabic morphs.

category	count	parameter a	parameter b
mono-syllabic	58	83.125	-0.652
di-syllabic	479	72.495	-0.609
tri-syllabic	127	7.463	-0.502





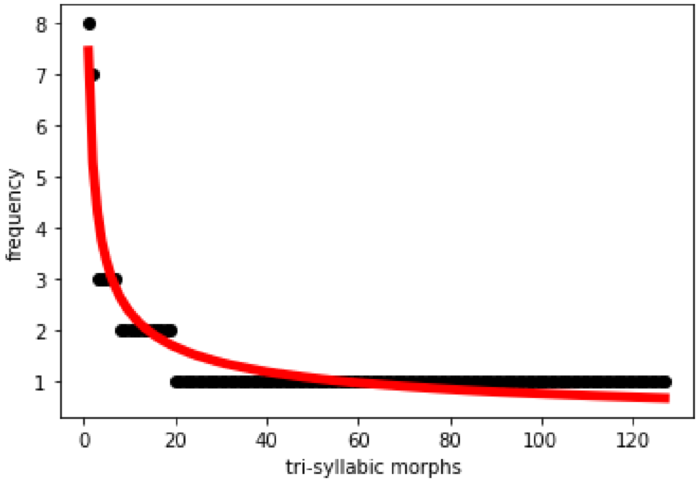


Figure 3. The distribution of mono, di, tri syllabic morphs is shown with the black dots in the three graphs correspondingly. The red curves in each graph show the curve fit using the power law parameters obtained for each of the distributions.

B. **Parameters at the syllable level:** The next step involved calculating parameters at the syllable level. Similar to the morph distributions, we model the distribution of type frequencies of all the possible syllables across morphs. The distribution of syllables was calculated from the bag of morphs, containing all distinct morphs in the data (morph types), and modeled using power law. The parameters obtained are a = 107.675, b = -0.589 . Figure 4 below shows the fitting of the power law over the syllable distribution.

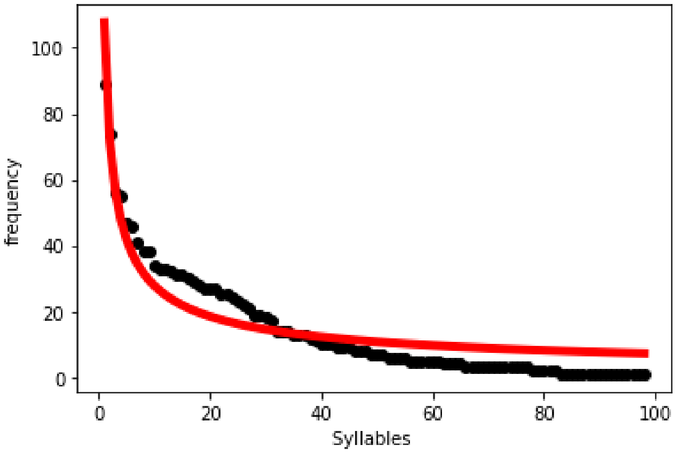


Figure 4. The distribution of syllable frequencies is shown by the black dots. The red curve shows the curve fit using the power law parameters obtained.

2. Generation of pseudo-Māori polymoraic words: Using the parameters calculated above, the pseudo Māori words were generated step-by-step through a bottom up generative process as explained below:

Step 1. Initialization of Consonants and Vowels : We start by defining the basic building blocks of the pseudo Māori language.This includes a set of consonants: ['h', 'f', 't', 'N', 'r', 'k', 'n', 'm', 'p', 'w'], and a set of vowels: ['a', 'e', 'i', 'o', 'u', 'A', 'E', 'I', 'O', 'U'], where ‘A’,’E’,’I’,’O’,’U’ are the long vowels and ‘wh’ is represented as f and ‘ng’ as ‘N’.

Step 2. Construct syllables : The syllable structures allowed in the language are ‘CV’ and ‘V’. Using this, all possible combinations following these syllable structures were created to create the bag of pseudo-syllables.

Step 3: Assigning Probabilities to Syllables : Using the power law parameter for syllables obtained above, np.random.power was used to assign probabilities to these pseudo-syllables. This step ensures that the frequency of syllable occurrence in the pseudo language mimics natural linguistic patterns.

Step 4. Generation of Syllabic Morphs : The number of mono, di and tri syllabic morphs were matched with category counts from the 1294 Māori words. The mono, di and tri syllabic morphs were generated from the bag of syllables using a sampling process (with replacement) with the syllable probabilities as weights.

Step 5. Probabilistic Weighting of Morphs : Once we had the bags of mono, di and tri syllabic pseudo morphs, the power law parameters for the three categories of morph categories were used to generate the probabilities (weights) for pseudomorphs. This probabilistic weighting is crucial for the next step, where these morphs are used to construct pseudo words.

Step 6. Construction of Pseudo Māori Words :To generate each pseudo word for a real Māori word, we used the syllable count to pick morphs with specific syllable counts by sampling (without replacement) from the respective bag of morphs. This ensured that we matched each Māori word with a statistically similar pseudo word. The selected morphs are then combined to form a pseudo word that statistically mirrors the syllable and morph structure of the real Māori word.

Step 6. Simulation and Analysis : We ran the generative process of generating words for 1000 iterations. In each iteration, a set of pseudo Māori words using the above method. After each iteration, the metrics were calculated across each language by averaging the word level metrics.

C. Results

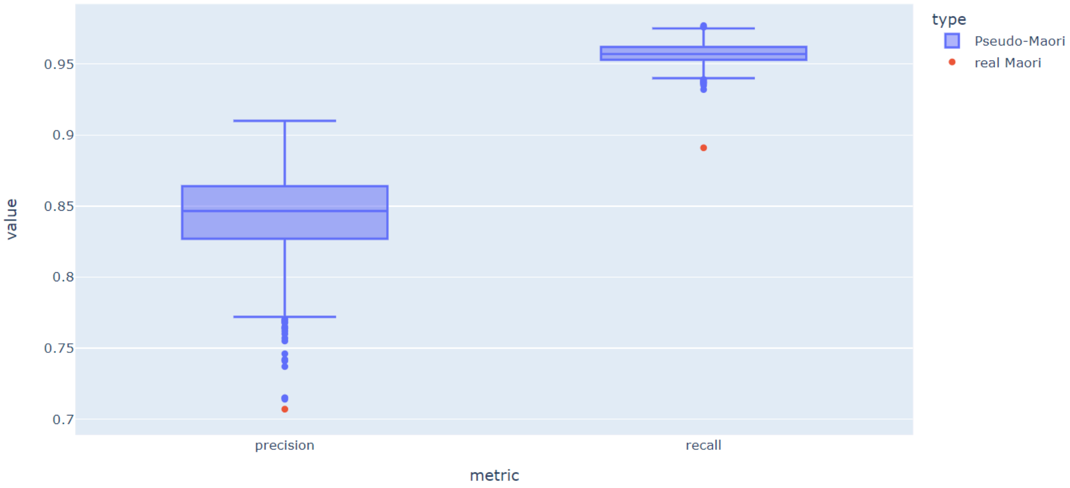


Figure 5. Box plot of Morfessor's precision and recall on the segmentations of 1000 sets of Pseudo-Māori polymoraic words (containing mono-syllabic, di-syllabic and tri-syllabic morphs only) along with Morfessor's performance on these subset of words from real Māori.

Figure 5 shows the distribution of macro-averaged precision recall (calculated mean for each language) for Morfessor's segmentations on all the 1000 generated Pseudo-Māori languages. The precision and recall were calculated at the word level, and then averaged across all the words to calculate the mean for each language which is the value used for the plots. With a mean precision of 0.843 and a mean recall of 0.957, we can see that Morfessor is able to segment pseudo-Māori really well. Since each word was matched with a real Māori word (excluding words containing 4 and 5 syllable morphs), we can calculate the precision and recall for these subset of words in real Māori. The mean precision and recall for these words from real Māori are 0.707 and 0.891.

The high precision and recall rates on pseudo-Māori polymoraic words indicate that Morfessor excels at identifying and segmenting patterns that are statistically present in the dataset. In pseudo-Māori , with its highly controlled design, the words are matched to the original words in terms of structure and regularity. The difference between the performance could arise from the fact that the gold standard segmentations of the real Māori words is a function not just of the morphological statistics, but also of other cues that the fluent speaker who provided the gold standard segmentations picked up on. By contrast, the pseudo-Māori words are based just on the morphological statistics which Morfessor can pick up on.

It could be the case that NMS also struggle with similar structural challenges, i.e. for the most part they are able to learn the statistical properties which occur in the polymoraic words, which can be seen by the high precision and recall by NMS in the polymoraic category. However, like Morfessor, they could be missing out on cues which they don't have access to. Further analysis is required to understand what are the exact kind of cues they struggle with.

VI. Discussion and Future work

By comparing Morfessor and NMS speakers in Analysis 1, we note how similar they performed in some word categories and how different they performed in others. The reduplication word category results show that NMS with a better precision are able to pick up these recurring patterns better than Morfessor. Reduplication is a typologically common feature, found in 85% of languages documented in the World Atlas of Language Structures (Rubino. (2013)), but it is often overlooked in unsupervised morphological segmentation approaches (Todd et al. (2022)). Incorporating reduplication templates directly into Morfessor was shown to substantially improve segmentations compared to the original Morfessor model (without reduplication templates) (Todd et al. (2022)). By using these templates in future analysis, we could analyze if Morfessor performance can be improved and observe if that helps Morfessor behave more like the NMS. If using reduplication templates in Morfessor helps improve the performance and mimic NMS, it could further help us gain proof of how NMS use ‘templatic’ approaches towards building their bag of morphs.

Within the polymoraic category of real Māori words, though these words closely match the underlying assumptions of Morfessor (these words are constructed through simple concatenation of morphs, without things like allomorphy or reduplication templates), NMS still outperformed Morfessor significantly. This shows that the implicit learning mechanism of NMS is more complex than that of Morfessor’s. As seen in analysis 1, NMS seem to follow a 'templatic' procedure in their learning process. They look for overarching templates in the language, which guide their segmentation. And these templates require a certain level of confidence before they start using them as rules in the language. In the affix analysis above, the difference in performances in the default and the non-default allmorphs is an example of this templatic approach. Their approach seems to be grounded in a broader understanding of the language's structure, where they wait to reach a certain threshold of understanding before finalizing the templatic knowledge they are acquiring. This behavior was noted in Panther et al.'s study, where it was highly unlikely for NMS to segment at positions other than the initial bimora. This also causes them to be overly cautious in placing boundaries, thus leading to under segmentation errors.

Zooming in on the errors made by Morfessor and NMS, it was interesting to see that both of them made different kinds of errors. While Morfessor's errors were mainly over segmentation, NMS were mostly undersegmenting. Like discussed above, this could be due to Morfessor's oversimplification of the language structure while NMS have not fully acquired the entirety of the sub parts of the words in the language or haven't reached the threshold yet to learn that these word parts can occur on their own. While they seem to be picking up on patterns which Morfessor misses (such as the reduplication patterns), they still lack in other parts of their morphological knowledge. Further investigation into what they lack could provide insights into their learning processes.

In Analysis 2, we saw that Morfessor is able to perform well on the pseudo-Māori compared to the real Māori words. This shows that the Morfessor indeed is good at picking up the statistical regularities in pseudo-Māori language. However, the reason it underperforms in real Māori could be due to information it does not have access to real Māori words – the cues which humans could be picking up on that could have led to the difference in performance. The following list could be some of the factors that contribute to these differences :

1. **Moraic Weights:** Moraic weight pertains to the role of moras in determining syllable structure and stress patterns. A mora is a basic timing unit in the phonology of some spoken languages, equal to or shorter than a syllable. The concept of moraic weight in Māori is closely tied to vowel length. Long vowels have more weight because they occupy two moras, influencing how syllables are perceived and stressed. At the moment, our language generation depends primarily on syllable weights, while the role of moraic weights might be significant. Moraic weights determine the length and rhythm of syllables and can have a profound impact on the prosody of a language. Humans are particularly sensitive to long vowels and diphthongs (Panther et al., 2023b), and our current approach may not fully account for these complexities.

2. **Affixes:** Our current system doesn't explicitly distinguish between affixes and root words. Incorporating these differences into the generator could help us understand if they perform better in one category over another. NMS again could be using an affix template approach where they identify there are certain morphs which can recur in certain positions with other morphs. By using a similar templatic approach as in Todd et al. (2022), we could provide information to Morfessor about affix templates to see if this helps Morfessor in its performance.

3. **Vowel and Consonant Harmony:** Presently, we're not differentiating between vowel harmony and consonant harmony. These phonological processes, where certain features of a sound spread to adjacent sounds, are key components in many

languages. Todd et al. (2021) show that in Māori, morphs generally show a preference for harmonic vowels and disharmonic constants. However, in compounds, vowel harmony is often not observed, which contrasts with previous findings that suggested a general tendency towards vowel harmony in words, thus illustrating the impact of phonological factors on morphology. By analyzing the segmentations for these properties, we can get insight into how it could affect the segmentations. Having the pseudo-Māori language have these properties can further provide insights into how NMS and Morfessor differ on these factors.

4. **Phonotactic Probabilities**: We are also not considering phonotactic probabilities, which refer to the likelihood of certain sounds occurring together in a language. Incorporating this could make our language generation more accurate and natural, as it would respect the rules governing the arrangement of phonemes within a language. Given that there have been unsupervised word segmentation models based on phonotactic probabilities (Daland and Pierrehumbert 2011), it is crucial to test how this affects morphological segmentation.

From this initial study at the morph level, we can infer that statistical cues in the language structure helps gain a certain level of morph identification in real Māori. However, further research has to be done to incorporate potential cues such as templatic approaches for different word categories, phonotactic probabilities, moraic weights etc into the generated languages to understand how Morfessor can gain Māori speaker level performance, thus providing insights into the exact regularities and patterns picked up by the Non-Māori speakers.

While Morfessor takes a Bayesian approach to morphological segmentation using the Minimum Description Length principle, using other segmentation models such as DiBS (Diphone-Based Segmentation) (Daland and Pierrehumbert 2011), could provide an understanding of how statistical properties in phonotactic features could help towards implicit language learning. DiBS analyzes how these features interact in a language, focusing on the segmentation of diphones (pairs of adjacent phonetic units) to understand the language's structure. As seen how providing reduplication template information to Morfessor helps its performance in the segmentation of reduplication word category (Todd et al. 2022), identifying ways to incorporate prosodic and phonotactic patterns can help provide a more holistic morphological segmentation model which could be closer to processes used by NMS.

More broadly, this work contributes to the understanding of statistical learning in humans and the cognitive implications by comparing non-speakers of a language who have ambient exposure to a morphological segmentation model. Both Morfessor and NMS seem to be using statistical learning; however, the patterns or regularities of the language that they learn appears to be different. NMS seem to understand patterns which Morfessor misses. This could be attributed to two factors as observed from our analysis. First, NMS seem to be using a broad pattern matching process using a templatic approach in order to draw their conclusions on which word parts are morphs as seen in the reduplication word category. Second, they potentially use a higher threshold than Morfessor to identify confidently if the word parts are indeed morphs. Further work is needed to tease apart these factors and confirm this hypothesis. These cues could further help us understand how unsupervised morphological segmentation models can be made better in order to pick up on these pattern identifying processes in addition to the statistical recurrence of morphs.

These results also help us provide insights into the role of language structure in unsupervised learning morphological segmentation. As we have seen here, it is important to know the strengths of how learning models perform in real languages. For example, from analysis 2 we did see that although Morfessor is good at picking statistical morph structure cues in highly constrained pseudo Māori, one needs to understand why it is not performing well in real Māori. Similar to how (Todd et al. 2022) showed that having reduplication templates could inform Morfessor better about reduplication patterns in the language, similar added cues to Morfessor and other unsupervised segmentation models can help learn them learn the morphological structural challenges of real languages.

One of the research questions of this work was to understand the extent of statistical learning of morphs in unsupervised learning through exposure in a non-lab setup, by morphological segmentation models and in implicit learning contexts with respect to Non-Māori speakers, both of whom share many properties in their learning processes. We have seen through our analysis the similarities in the statistical properties which both Morfessor and NMS learn, and connecting it to the cues which are available to the two learners. NMS, as we saw, have access to external cues such as phonotactics and the cognitive mechanism of template approach, which need further analysis to understand their impact completely.

Based on the comparisons between Morfessor's performance on real Māori and pseudo Māori words, we were able to address our second research question. From our analysis, we saw that Morfessor is able to segment the pseudo-Māori words since they follow the concatenative properties. This helped us point out that it relatively struggles with words in real Māori since there are more factors to the real words which need other kinds of cues which we need to understand further. Since NMS are able to perform better, we can say that NMS seem to be accessing these cues which Morfessor doesn't have access to. Further analysis is required to identify and incorporate these cues into Morfessor to check if it helps improve Morfessor's performance, thus proving our hypothesis.

Lastly, this work also helped point out the ways in which unsupervised segmentation models like Morfessor need additional language-specific aids for segmentation. The difference between performances by Morfessor on pseudo Māori and real Māori words underscores the importance of understanding the differences in morphological segmentation model output when applied to different languages. It helps us think about how these models can be made better by providing additional cues to it as shown by the reduplication templates in Todd et al. (2022).

This work has helped dive into understanding of what underlying cognitive processes the NMS could be using by doing an analysis on the segmentations by the two learner models. It provides us with further directions to build a complete understanding of the processes underlying the implicit statistical learning by NMS.

VII. Conclusion

In summary, our comparative analysis of Morfessor and NMS in handling Māori word categories underscores the nuanced complexities in language processing. While Morfessor's algorithmic precision captures certain statistical regularities, it falls short in the intricate aspects of linguistic structure that NMS, through implicit learning mechanisms, navigate more adeptly. This divergence not only highlights the limitations of current unsupervised morphological segmentation models but also points to potentially rich, templatic patterns employed by NMS. Understanding the underlying processes of NMS could pave the way for enhancing computational models by integrating more sophisticated, human-like implicit learning processes.

Furthermore, this study's findings significantly contribute to our comprehension of statistical learning in language acquisition in the context of implicit learning. The distinct approaches employed by NMS and Morfessor in deciphering linguistic patterns reveal the depth and variety of statistical properties that inform language understanding. The superior performance of NMS in certain word categories suggests a threshold-based, templatic learning process, which is less evident in Morfessor’s method. Future research on dissecting these learning mechanisms, particularly how they apply to unsupervised language learning scenarios could help improve unsupervised models when applied to languages such as Māori, which are not commonly used in benchmarking these models. The insights from such investigations could improve our approaches to computational language models, making them more reflective of the intricate, multi-layered processes as seen in the implicit learning process by NMS. In essence, exploring these avenues could lead to more human-like models, greatly enhancing our ability to model and understand the complexities of morphological segmentation specifically and implicit learning more broadly.

References

Bulf, Hermann, Scott P. Johnson, and Eloisa Valenza. 2011. “Visual Statistical Learning in the Newborn Infant.” *Cognition* 121 (1): 127–32. https://doi.org/10.1016/j.cognition.2011.06.010.

Creutz, Mathias, and Krista Lagus. 2007. “Unsupervised Models for Morpheme Segmentation and Morphology Learning.” *ACM Transactions on Speech and Language Processing* 4 (1): 3:1-3:34. https://doi.org/10.1145/1187415.1187418.

———. n.d. “Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.”

Daland, Robert, and Janet B. Pierrehumbert. 2011. “Learning Diphone-Based Segmentation.” *Cognitive Science* 35 (1): 119–55. https://doi.org/10.1111/j.1551-6709.2010.01160.x.

Estes, Katharine Graf, Julia L. Evans, Martha W. Alibali, and Jenny R. Saffran. 2007. “Can Infants Map Meaning to Newly Segmented Words?: Statistical Segmentation and Word Learning.” *Psychological Science* 18 (3): 254–60. https://doi.org/10.1111/j.1467-9280.2007.01885.x.

Frank, Michael C., Joshua B. Tenenbaum, and Edward Gibson. 2013. “Learning and Long-Term Retention of Large-Scale Artificial Languages.” *PLOS ONE* 8 (1): e52500. https://doi.org/10.1371/journal.pone.0052500.

Gervain, Judit, Francesco Macagno, Silvia Cogoi, Marcela Peña, and Jacques Mehler. 2008. “The Neonate Brain Detects Speech Structure.” *Proceedings of the National Academy of Sciences* 105 (37): 14222–27. https://doi.org/10.1073/pnas.0806530105.

Johnson, Elizabeth K. 2016. “Constructing a Proto-Lexicon: An Integrative View of Infant Language Development.” *Annual Review of Linguistics* 2 (1): 391–412. https://doi.org/10.1146/annurev-linguistics-011415-040616.

Ngon, Céline, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat, and Sharon Peperkamp. 2013. “(Non)Words, (Non)Words, (Non)Words: Evidence for a Protolexicon during the First Year of Life.” *Developmental Science* 16 (1): 24–34. https://doi.org/10.1111/j.1467-7687.2012.01189.x.

Oh, Y., S. Todd, C. Beckner, J. Hay, and J. King. 2020. “Non-Māori-Speaking New Zealanders Have a Māori Proto-Lexicon.” *Scientific Reports* 10 (1): 22318. https://doi.org/10.1038/s41598-020-78810-4.

Oh, Y., S. Todd, C. Beckner, J. Hay, J. King, and J. Needle. 2020. “Non-Māori-Speaking New Zealanders Have a Māori Proto-Lexicon.” *Scientific Reports* 10 (1): 22318. https://doi.org/10.1038/s41598-020-78810-4.

Panther, Forrest Andrew, Wakayo Mattingley, Simon Todd, Jennifer Hay, and Jeanette King.

2023. “Proto-Lexicon Size and Phonotactic Knowledge Are Linked in Non-Māori Speaking New Zealand Adults.” *Laboratory Phonology* 14 (1). https://doi.org/10.16995/labphon.7943.

Panther, Forrest, Wakayo Mattingley, Jen Hay, Simon Todd, Jeanette King, and Peter J. Keegan. 2023. “Morphological Segmentations of Non-Māori Speaking New Zealanders Match Proficient Speakers.” *Bilingualism: Language and Cognition*, June, 1–15. https://doi.org/10.1017/S1366728923000329.

Pelucchi, Bruna, Jessica F. Hay, and Jenny R. Saffran. 2009. “Statistical Learning in a Natural Language by 8-Month-Old Infants.” *Child Development* 80 (3): 674–85. https://doi.org/10.1111/j.1467-8624.2009.01290.x.

Rebuschat, Patrick. 2015. *Implicit and Explicit Learning of Languages*. John Benjamins Publishing Company.

Rissanen, J. 1978. “Modeling by Shortest Data Description.” *Automatica* 14 (5): 465–71. https://doi.org/10.1016/0005-1098(78)90005-5.

Saffran, Jenny R. 2001. “Words in a Sea of Sounds: The Output of Infant Statistical Learning.” *Cognition* 81 (2): 149–69. https://doi.org/10.1016/S0010-0277(01)00132-9.

Saffran, Jenny R. 2003. “Statistical Language Learning: Mechanisms and Constraints.” *Current Directions in Psychological Science* 12 (4): 110–14. https://doi.org/10.1111/1467-8721.01243.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. 1996. “Statistical Learning by 8-Month-Old Infants.” *Science* 274 (5294): 1926–28. https://doi.org/10.1126/science.274.5294.1926.

Saffran, Jenny R., Elissa L. Newport, Richard N. Aslin, Rachel A. Tunick, and Sandra Barrueco. 1997. “Incidental Language Learning: Listening (and Learning) Out of the Corner of Your Ear.” *Psychological Science* 8 (2): 101–5. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x.

Smit, Peter, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. “Morfessor 2.0: Toolkit for Statistical Morphological Segmentation.” https://aaltodoc.aalto.fi:443/handle/123456789/14051.

Stolcke, A., and E. Shriberg. 1996. “Automatic Linguistic Segmentation of Conversational Speech.” In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96*, 2:1005–8 vol.2. https://doi.org/10.1109/ICSLP.1996.607773.

Teinonen, Tuomas, Vineta Fellman, Risto Näätänen, Paavo Alku, and Minna Huotilainen. 2009. “Statistical Language Learning in Neonates Revealed by Event-Related Brain Potentials.” *BMC Neuroscience* 10 (1): 21. https://doi.org/10.1186/1471-2202-10-21.

Todd, Simon, Annie Huang, Jeremy Needle, Jennifer Hay, and Jeanette King. 2022. “Unsupervised Morphological Segmentation in a Language with Reduplication.” In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, edited by Garrett Nicolai and Eleanor Chodroff, 12–22. Seattle, Washington: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.2.

Todd, Simon, Chadi Ben Youssef, and Alonso Vásquez-Aguilar. 2023. “Language Structure, Attitudes, and Learning from Ambient Exposure: Lexical and Phonotactic Knowledge of Spanish among Non-Spanish-Speaking Californians and Texans.” *PLOS ONE* 18 (4): e0284919. https://doi.org/10.1371/journal.pone.0284919.

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*. Aalto University. https://aaltodoc.aalto.fi:443/handle/123456789/11836.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Williams, John N. 2020. "The Neuroscience of Implicit Learning." *Language Learning* 70 (S2): 255–307. https://doi.org/10.1111/lang.12405.

Moorfield, J. C. Te Aka: Māori-English, English-Māori dictionary and index. Pearson, Auckland (2011). (third edition).

Rubino, C. (2013). Reduplication. The World Atlas of Language Structures Online.

Harlow, Ray. 2007. Māori: A linguistic introduction. Cambridge: Cambridge University Press.

Simon Todd, Jeremy Needle, Jennifer Hay & Jeanette King (2021). Phonological Influences on Lexicalized Compound Formation in Māori. Talk presented at 95th Annual Meeting of the Linguistic Society of America.