More than Just Statistical Recurrence: Human and Machine Unsupervised Learning of Māori Word Segmentation across Morphological Processes

Ashvini Varatharaj
Department of Linguistics,
University of California Santa Barbara
ashvinivaratharaj@ucsb.edu

Simon Todd
Department of Linguistics,
University of California Santa Barbara
& NZILBB, University of Canterbury
sjtodd@ucsb.edu

Abstract

Non-Māori-speaking New Zealanders (NMS) are able to segment Māori words in a highly similar way to fluent speakers (Panther et al., 2024). This ability is assumed to derive through the identification and extraction of statistically recurrent forms. We examine this assumption by asking how NMS segmentations compare to those produced by Morfessor, an unsupervised machine learning model that operates based on statistical recurrence, across words formed by a variety of morphological processes. Both NMS and Morfessor succeed in segmenting words formed by concatenative processes (compounding and affixation without allomorphy), but NMS also succeed for words that invoke templates (reduplication and allomorphy) and other cues to morphological structure, implying that their learning process is sensitive to more than just statistical recurrence.

1 Introduction

Humans have a powerful ability to build implicit linguistic knowledge incidentally, based on passive processes that identify and extract statistically recurrent patterns (Saffran et al., 1996; Frank et al., 2013; Aslin, 2017). For example, New Zealanders who are regularly ambiently exposed to Māori, but do not speak it, nevertheless have Māori lexical and phonotactic knowledge (Oh et al., 2020; Panther et al., 2023) and can morphologically segment Māori words at above-chance levels (Panther et al., 2024). These findings imply that regular exposure to a language yields a *proto-lexicon*: an implicit memory-store of forms that recur with statistical regularity in the language, including both words and word-parts (Ngon et al., 2013; Johnson, 2016).

In this paper, we are concerned with the way that the proto-lexicon is constructed, and the way that its construction interacts with language structure. We examine the extent to which the ability of non-Māori-speaking New Zealanders (NMS) to morphologically segment Māori words is explained by naive statistical learning, in which their proto-lexicon is assumed to be formed purely through the identification and extraction of statistically recurrent forms in ambient Māori. To do so, we generate expectations for what morphological segmentation would look like through naive statistical learning processes from Morfessor (Creutz and Lagus, 2007; Virpioja et al., 2013), an unsupervised Bayesian segmentation model. We compare the segmentations produced by Morfessor to those produced by NMS and examine how they vary across words formed by different morphological processes.

Through two analyses, we argue that NMS do more than a naive statistical learning model would suggest. First, we compare the segmentations of Morfessor and NMS across Māori words formed by affixation and compounding, both concatenative processes, and words formed by reduplication, a templatic process. We find that both are accurate on words formed by affixation and compounding, but NMS are more accurate on words formed by reduplication, suggesting that NMS identify and extract both statistically recurrent forms and higher-level abstract templates. Then, zooming in on words formed by concatenative processes, we ask whether there are other cues to morphological structure that NMS may be picking up on, such as vowel length. We compare the performance of Morfessor across real Māori words that may contain such cues and constructed words that have the same statistical properties but lack any reliable alternative cues to morphological structure. We find that Morfessor is worse at segmenting real words, suggesting that successful learning by NMS requires sensitivity to more cues than just statistical recurrence.

2 Background

2.1 Statistical learning of language

How humans learn to extract knowledge from their environment is one of the fundamental questions in

Proceedings of the 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 20–31

cognitive science. Implicit learning – the process of learning without intention, and even without the awareness of what has been learned (Williams, 2020) – is one of the main ways we learn from our surroundings. Implicit learning 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*[1]. 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 et al., 2011; Gervain et al., 2008; Teinonen et al., 2009).

While most work on statistical learning has focused on studying infants (Saffran, 2001; Pelucchi et al., 2009) in lab-based setups, recent works have shown that adults are also capable of statistical learning of implicit linguistic knowledge through everyday exposure to a language they don't speak. Non-speakers of Māori in New Zealand (Oh et al., 2020; Panther et al., 2023) and Spanish in California and Texas (Todd et al., 2023) show evidence of implicit phonotactic and lexical knowledge of their respective ambient languages. However, this knowledge appears to be weaker in the case of Spanish than in Māori, and it has been argued that this difference may partly derive from differences in morphological structure (Todd et al., 2023).

In addition to having implicit phonotactic and lexical knowledge of Māori, non-Māori-speaking New Zealanders (NMS) can morphologically segment Māori words in a highly similar way to fluent speakers (Panther et al., 2024). This ability is facilitated by their possession of a *proto-lexicon* (Johnson, 2016; Ngon et al., 2013), a large implicit memory-store of the forms of words and word-parts that recur with statistical regularity in the language, called *morphs*. These morphs are defined by form, without consideration of meaning; thus, they may or may not correspond to underlying morphemes, and may even include phonological sequences that span word boundaries as long as they are statistically recurrent in the language (Ngon et al., 2013).

[1]While early literature on statistical learning focused narrowly on phonotactic transition probabilities, in this work we use the term more broadly to refer to the learning of any statistical properties of language.

2.2 Morphological segmentation

Many modern approaches to morphological segmentation use supervised learning, independently or in combination with unsupervised learning (e.g., Rouhe et al., 2022). In this work, we are attempting to model human learning of morphological segmentation that occurs without explicit instruction. For this reason, we use unsupervised learning.

Unsupervised morphological segmentation provides us an avenue to simulate implicit statistical learning processes. In this work, we use Morfessor Baseline (Creutz and Lagus, 2007; Virpioja et al., 2013), a popular unsupervised morphological segmentation model with an underlying generative process that is very simple and highly compatible with a naive model of statistical learning of morphological structure. Morfessor identifies a set of statistically recurrent morphs under the assumption that words are formed through the concatenation of these morphs, without phonological alternations, and without constraints applied to positioning, sequencing or morphosyntactic category.

Morfessor identifies the set of statistically recurrent morphs, which it calls a *lexicon* (and which is analogous to a human proto-lexicon), using a Minimum Description Length framework (Rissanen, 1978). This lexicon is therefore the smallest set of simplest morphs that can be combined to generate the training data with highest probability. The lexicon is constructed dynamically through several passes over the training data, where the cost of adding a morph to the lexicon at any point is based on the morph's complexity and its frequency of recurrence across the words segmented so far.

While Morfessor's assumptions are simple, there are simpler models that have gained currency recently as tokenizers in Natural Language Processing (e.g., Sennrich et al., 2015; Kudo, 2018; Wu et al., 2016). Like Morfessor, these models identify a set of morphs (which they call *subwords*) that generate the training data with highest probability, assuming only simple concatenation. However, unlike Morfessor, they require the number of morphs to be predetermined, and they do not simultaneously consider the complexity of proposed morphs, which we consider to be important for our modeling of human learning. There are also many morphological segmentation models that are more complex than Morfessor, such as Adaptor Grammars (Johnson and Griffiths, 2007; Eskander et al., 2016; Godard et al., 2018). These models offer fine-

grained assumptions about precisely how morphs may be combined, in contrast to Morfessor's assumption of simple concatenation. It is the relative simplicity of Morfessor that makes it a suitable baseline model of idealized statistical learning of a proto-lexicon, especially in a language that uses primarily concatenative morphological processes.

Morfessor's statistical learning approach mirrors that which has been assumed for NMS (Oh et al., 2020). Both are learning to segment based on statistical patterns in the language they are exposed to, without getting feedback. In both cases, the learners are identifying recurring forms and extracting them as morphs in a (proto-)lexicon. By using Morfessor as a baseline of comparison for NMS, we can understand how much of NMS' implicit knowledge is due to simple statistical learning processes. We expect Morfessor to perform best with words formed by concatenative morphological processes and to struggle with words formed by other morphological processes that are beyond the scope of its simple assumptions; if NMS do not struggle in the same way, then we may infer that they are doing more than just tracking statistical recurrence as Morfessor would assume.

2.3 The Māori Language

The Māori language consists of ten consonants <p, t, k, m, n, ng, w, r, wh, h>, five short vowels <a, e, i, o, u>, and five long vowels <ā, ē, ī, ō, ū>. The orthographic system is highly transparent: each grapheme or digraph corresponds to a unique phoneme. The basic timing unit is the mora, where short vowels count as one mora each and long vowels count as two (Harlow, 2007). The syllable structure is (C)V(V), but is often treated as (C)V for modeling purposes because of the complexity of distinguishing diphthongs from sequences of monopthongs (Bauer, 1993; Oh et al., 2020). There is a general minimality constraint which states that (content) words and morphs consist of at least two moras (Bauer, 1993, p. 544), and it has been argued that words consisting of four or more moras are highly likely to be morphologically complex (Krupa, 1968; de Lacy, 2003).

There are three main morphological processes in Māori: reduplication, affixation, and compounding (Bauer, 1993; Harlow, 2007). Reduplication consists of the repetition of part of a base, following one of many templates (see e.g. Keegan, 1996; Todd et al., 2022). Because of this reliance on a template, we refer to reduplication as a *templatic* process.[2] Affixation and compounding both consist of the concatenation of morphs that need not have any relation to each other in form, and thus we refer to them as *concatenative* processes. At a distributional level, affixation and compounding are distinguished by the fact that affixation causes a small set of four (Bauer and Bauer, 2012) or five (Harlow, 2007) productive morphs[3] to recur across many words, whereas compounding causes a large set of morphs to each recur across relatively fewer words (Bauer, 1993, p. 519).

Māori morphophonology may be described as strictly local: there are no morphophonological alternations, no phonologically discontiguous morphemes, and no long-distance phonological dependencies. However, there is affix allomorphy, in which affixes follow phonological templates, with different thematic consonants that are to some extent predictable (Parker Jones, 2008). This allomorphy is restricted to the passive and nominalizing suffixes, each of which has default and non-default allomorphs that are or are not consistent with major phonological templates (passive: *-Cia*; nominal: *-Canga*; both for thematic consonant *C*).

At a high level, the strictly local nature of Māori morphophonology accords exactly with the assumptions of Morfessor. However, the templates that underpin reduplication and affix allomorphy are not accounted for by Morfessor's underlying generative model. This means that the three morphological processes in Māori are consistent with Morfessor's assumptions to different extents, which allows us to examine how the degree to which Morfessor reflects NMS morphological segmentations is affected by morphological structure.

3 Analysis 1: Sensitivity to templates

Our first analysis examines Morfessor and NMS segmentations of Māori words formed through different morphological processes. For each learner, we identify the sensitivity to general templates and the importance of morphological concatenativity by comparing segmentation performance across words formed by reduplication and words formed by affix-

[2] We avoid the label templatic *morphology* so as to avoid confusion with root-and-pattern morphology such as is found in Semitic languages.

[3] Whether there are held to be four or five productive affixes depends on where the analyst draws the line between affixation and phrasal constructions. It is not entirely straightforward to designate these affixes as clearly inflectional or clearly derivational (Bauer and Bauer, 2012).

ation or compounding (Section 3.2), as well across cases of affixation that follow salient allomorphic templates to different extents (Section 3.3). This analysis reveals how unsupervised learning of morphological segmentation is sensitive to linguistic structure, and the extent to which the underlying assumptions of Morfessor make it a plausible model of naive statistical learning of morphological segmentation in humans.

3.1 Data

The analysis is conducted over a subset of words from the stimuli of Panther et al. (2024), which we aggregated into categories based on the morphological processes they likely represent (described below). We used the segmentations provided by a fluent Māori speaker (MS), collected by Oh et al. (2020), as a gold standard. To ensure that the morphological processes assumed by our categorizations adequately reflect those revealed by the MS segmentations, we filtered each category to only include words in which the MS segmentation is consistent with the assumed morphological process. After this filtering, the analysis is based on 3,919 words, categorized as follows:

Monomorphemic: Words consisting of 2 or 3 moras ($N = 622$ / 295, respectively) that did not receive any boundaries in the MS segmentation.

Reduplication: Words that were segmented by the MS in a manner consistent with one of four reduplication templates[4]: total (e.g., *paki+paki*; $N = 439$), right (e.g., *tākai+kai*; $N = 276$), left (e.g., *nu+nui*; $N = 111$), or left with lengthening (e.g., *kā+kahu*; $N = 36$). Total reduplication is the most salient of these templates.

Affixation: Words in which the MS recognized either the causative prefix *whaka-* ($N = 296$), a passive suffix ($N = 437$), or a nominalizing suffix ($N = 203$). The suffixes have many allomorphs which differ in terms of frequency and consistency with a major phonological template (passive: *-Cia*; nominal: *-Canga*; both for thematic consonant *C*), including, in descending order of frequency: template-consistent defaults (passive: *-tia*, *-hia*, *ngia*; nominal: *-tanga*, *-hanga*)[5]; non-template-consistent defaults (passive: *-a*; nominal: *-nga*[6]); template-consistent non-defaults (passive: *-kia*, *-mia*, *-ria*, *-whia*; nominal: *-kanga*, *-manga*, *-ranga*, *-whanga*); and non-template-consistent non-defaults (passive: *-ia*, *-na*, *-nga*[6], *-ina*, *-hina*, *-kina*, *-whina*; nominal: *-anga*).

Compounding: Words that consist of four or more moras, without reduplication or affixation, and for which the MS identified at least one boundary ($N = 1204$; a subset of the 'polymoraics' explored by Panther et al., 2024).

For each word, we compare the gold standard segmentation provided by the MS to the segmentations provided by Morfessor and NMS. The Morfessor segmentations were obtained from a model trained with default settings (using the implementation of Virpioja et al., 2013) on 19,595 word types from the Te Aka dictionary (Moorfield, 2011). The NMS segmentations are based on data collected by Panther et al. (2024) in a word-splitting task, where NMS participants split orthographically-presented words into pieces by placing any number of boundaries at any site between two letters.[7] To aggregate segmentations of a single word across participants, we used a majority-vote approach: we coded each site as containing a boundary if and only if the majority of participants who responded to that word placed a boundary there.

3.2 Analysis 1A: Morphological processes

We first analyze the degree to which segmentations by Morfessor and NMS match the gold standard segmentations, across categories of words formed by different morphological processes. We examine variation across categories, as well as how this variation differs between learners.

3.2.1 Methods

There are many metrics that compare a learner's morphological segmentations to a gold standard (Virpioja et al., 2011). We use the simple metric of boundary precision and recall, which considers

[4]The reduplication category includes some cases where there is both reduplication and compounding. We assess the placement of all boundaries in such cases, regardless of whether they separate the reduplicant from the base or one compound component from another.

[5]Dialects differ in terms of whether the default thematic consonant is <t>, <h>, or <ng>, though it is most commonly <t> (Harlow, 2007).

[6]*-nga* is both a passive suffix and a nominalizing suffix. As a passive suffix, it is not a default allomorph, but as a nominalizing suffix, it is. Our analysis of *-nga* is restricted to its occurrence as a nominalizing suffix.

[7]We analyze the same filtered subset of NMS participants as Panther et al. (2024): 195 individuals who have lived in NZ since the age of 7, have never taken any linguistics courses, and have explicit knowledge of few Māori words and grammatical structures. For full details of the experiment design and filtering criteria, see Panther et al. (2024).

Table 1: Macro-averaged precision and recall for Morfessor and NMS across categories of words formed by different morphological processes.

	Morfessor		NMS	
Category	Prec.	Rec.	Prec.	Rec.
monomorphemic	0.66	0.66	0.79	0.79
reduplication	0.58	0.51	0.85	0.86
affixation	0.92	0.90	0.70	0.70
compounding	0.88	0.91	0.84	0.84

each potential boundary site independently. Precision in this context refers to the proportion of the sites identified by the learner as containing a boundary that also contain a boundary in the gold standard segmentation. Recall refers to the proportion of the sites containing a boundary in the gold standard segmentation that are identified by the learner as containing a boundary. We take a macro-averaging approach: we calculate precision and recall separately for each word, then average each metric across all words in each category. If precision and recall are both undefined for a word (i.e., if the gold standard segmentation contains no boundaries and the learner does not identify any), we set them both to 1; if only one metric is undefined, we set that metric to 0.

3.2.2 Results

The macro-averaged precision and recall for Morfessor and NMS across the four categories of words are shown in Table 1.

For monomorphemic words, both learners show indications of oversegmentation, via low precision and recall that result from placing boundaries where they shouldn't exist. NMS appear to show less oversegmentation than Morfessor, suggesting that they may be more sensitive to word minimality constraints based on moraic weight (Bauer, 1993, p. 544). This tendency toward oversegmentation does not stand out for either learner across other categories: precision and recall are fairly balanced for both learners across all categories, indicating a general balance between oversegmentation and undersegmentation.

For words formed by reduplication, a templatic process, NMS show better performance than Morfessor. This difference is made even clearer when considering performance on reduplication in relation to affixation and compounding (concatenative processes): for Morfessor, performance on reduplication is notably worse than performance on affixation and compounding, but for NMS, it is not. This result suggests that NMS may be sensitive to abstract reduplication templates that Morfessor cannot capture (Todd et al., 2022), and thus that their recognition of such templates may boost implicit learning above and beyond that expected from simple statistical learning of recurrent forms. In support of this suggestion, we found that Morfessor has worst performance on the subset of words formed by total reduplication, the most salient reduplication template, whereas NMS has best performance on this subset (precision/recall for Morfessor: 0.35/0.36; for NMS: 0.95/0.97).

For words formed by affixation and compounding, both concatenative processes, Morfessor performs well, suggesting that such words facilitate implicit learning of morphs via naive statistical learning. Nevertheless, it is somewhat surprising that Morfessor did not perform even better for these words, given that they exactly match the assumptions of its underlying generative model. This suggests that the morphological structure of Māori, as captured by the gold standard segmentations, may be cued by more than just the statistical recurrence of forms (Todd et al., 2019; Panther et al., 2024); we return to this point in Analysis 2 (Section 4).

NMS perform slightly worse than Morfessor on words formed by compounding, and notably worse on words formed by affixation. One possible interpretation of this result is that NMS are not as good at tracking statistical recurrence as Morfessor – hence the worse performance on both categories – but make up for this shortcoming to some extent in compounds by being sensitive to additional cues to morphological structure (Panther et al., 2024). The fact that NMS' difficulties are concentrated in words formed by affixation suggests that they may struggle specifically with recognizing affixes as independent of stems. A finer-grained inspection suggests that this may be related to issues of affix position, allomorphy, and/or frequency: NMS perform as well as Morfessor on words containing the highly frequent causative prefix *whaka-* (precision/recall for Morfessor: 0.95/0.93; for NMS: 0.95/0.93), which has no allomorphs, but perform worse on words containing passive or nominalizing suffixes (precision/recall for Morfessor: 0.90/0.89; for NMS: 0.59/0.59), which have many allomorphs, including some that are quite infrequent.

3.3 Analysis 1B: Affix recovery

To dig further into potential sources of issues with segmenting words formed by affixation, we analyze the ability of Morfessor and NMS to recover different affixes by segmenting them off. This analysis separates the causative prefix from passive and nominalizing suffixes, and subdivides passive and nominalizing allomorphs into smaller groups.

3.3.1 Methods

The affixes we analyze are organized into groups based on word position, status as default/non-default allomorph, and consistency with a major phonological template. The groups also vary in frequency. We define the type frequency of an affix group as the proportion of the 19,595 words for which Oh et al.'s (2020) MS segmented off a morph with the same form as some affix in the group, at the appropriate word edge. We similarly define the token frequency of an affix group as the proportion of tokens in the MAONZE corpus (King et al., 2011) and the Māori Broadcast Corpus (Boyce, 2006) that correspond to words for which the MS separated off some affix from the group.[8] Type frequency is relevant for Morfessor, and both type and token frequency may be relevant for NMS.

For each affix group, we measure the rate at which Morfessor and NMS successfully recover affixes in that group by segmenting them off words. We assign each word in the affixation category to one or more groups based on the affix(es) in its gold standard segmentation. For a word in a given group, a learner successfully recovers the affix pertaining to that group if their segmentation contains a boundary at the site between the affix and the rest of the word, without also containing any boundaries at sites within the affix. The segmentation of the stem is irrelevant: a learner can successfully recover an affix from a word even if their segmentation of the rest of the word does not match that represented by the gold standard segmentation. We measure the rate of affix recovery for a group as the proportion of words in the group for which the affix is successfully recovered.

3.3.2 Results

The affix recovery rates for each learner across the various affix groups are shown in Table 2.

[8]We follow Oh et al. (2020) in using Simple Good-Turing smoothing (Gale and Sampson, 1995) to ensure that words from the dictionary that were not mentioned in the corpora have a non-zero token frequency.

Both Morfessor and NMS have extremely high recovery rates for *whaka-*. This is not surprising, as it is extremely frequent, in terms of both types and tokens. For NMS, it is also highly salient due to its position at the beginning of verbs that often appear utterance-initially as imperatives (e.g., *whakarongo mai!* 'listen!') and its appearance in place names (e.g., Whakatane) and the well known and highly culturally significant word *whakapapa* 'genealogy' (Oh et al., 2023). There is also reason to believe that NMS may be particularly sensitive to prefixes such as *whaka-* because they have been shown to apply a bimoraic template when segmenting the first morph in a word (Panther et al., 2024).

While Morfessor and NMS have near-identical rates for the causative prefix *whaka-*, their recovery rates for allomorphs of the passive and nominalizing suffixes diverge, with NMS being less successful than Morfessor. One possible reason for this divergence is that NMS may be less sensitive to suffixes than prefixes, since the bimoraic template that facilitates sensitivity to prefixes operates from left to right and thus may not consistently align with suffixes. Another possible reason may stem from NMS being sensitive to token frequency rather than just type frequency like Morfessor, as the passive and nominalizing suffixes have much lower token frequencies than type frequencies, both in absolute terms and in relation to the corresponding frequency of *whaka-*. From a statistical learning perspective, a morph needs to be experienced sufficiently often in a range of environments before a learner can reliably identify and extract it, so lower experiential frequency by NMS compared to Morfessor would yield noisier segmentations. This difference is magnified by the fact that Morfessor has perfect memory of all types it has encountered at each point of the learning process, which is not the case for NMS.

Zooming into the allomorphs, for Morfessor there is a clear separation between default and non-default. This separation is driven by frequency: allomorphs in a given affix group can be recovered reliably if and only if they recur across sufficiently many types. After adjusting for frequency, there are no major differences between affix groups based on consistency with a major phonological template, nor phonological shape generally (e.g., passive CVV vs. nominalizing CVCV: recovery rates 0.964 and 0.966, respectively). This is not surprising: Morfessor's naive statistical learning

Table 2: Affix recovery rates for Morfessor and NMS across different affix groups. Affix groups vary in terms of position, status of allomorphs as default/non-default, consistency of allomorphs with major phonological templates, and frequency of occurrence (proportion of types / tokens affixed by that form).

		Frequency		Affix recovery	
Affix(es)	Allomorph group	type	token	Morf.	NMS
whaka-	–	0.142	0.017	0.983	0.976
-tia, *-tanga*	default, template[5]	0.128	0.006	0.995	0.783
-hia, *-ngia*, *-hanga*	default, template[5]	0.064	0.005	0.995	0.688
-a, *-nga*[6]	default, non-template	0.034	0.011	0.907	0.293
-kia, *-mia*, *-ria*, *-whia*, *-kanga*, *-manga*, *-ranga*	non-default, template	0.017	0.003	0.702	0.553
-ia, *-na*, *-ina*, *-hina*, *-kina*, *-whina*, *-anga*	non-default, non-template	0.016	0.002	0.739	0.370

algorithm has no access to phonological templates, and is based primarily on frequency.

For NMS on the allomorphs, there is also a relationship between affix recovery rate and frequency, but it is more gradient, reflecting differences in the experiential frequency and memory of NMS compared to Morfessor. The correlation is not perfect, however. The affix recovery rate is extremely low for the default allomorphs that are not consistent with a template, in spite of their high token frequency. It is also higher than expected for the non-default allomorphs that are consistent with a major phonological template, in comparison to those that have almost identical frequency but are not consistent with a template. These results suggest that NMS are sensitive to major phonological templates, giving them an advantage in recognizing allomorphs that are consistent with them.

Furthermore, since the default allomorphs that are not consistent with a template are also short – with one simply having the shape V – the fact that NMS recover them less successfully suggests a sensitivity to phonological shape generally. That is, NMS may find morphs less salient the less phonological content they have and/or the less their syllables resemble the canonical CV shape. This suggestion is further supported by the fact that NMS are less successful at recovering passive suffixes with a CVV shape than nominalizing suffixes with a CVCV shape (rates 0.669 and 0.866, respectively).

4 Analysis 2: Other cues

Analysis 1 showed that NMS are sensitive to templates, both at the word level (reduplication) and at the morph level (minimality constraints; allomorphs that follow a phonological template or feature syllables with canonical CV shape). Morfessor shows no such sensitivity, as its underlying generative model does not incorporate templates, and thus underperforms when segmenting words that invoke templates in some way.

However, templates appear not to be the only reason that Morfessor underperforms. In Section 3.2.2, we observed that Morfessor's performance on compounds was lower than might be expected, given that they follow its underlying assumption of morphological concatenativity. Based on this observation, we suggested that the morphological structure of Māori may be cued by more than the statistical recurrence of forms, consistent with previous results showing that MS segmentations are sensitive to aspects such as the presence of long vowels (Todd et al., 2019; Panther et al., 2024).

Here, we explore this suggestion further by comparing Morfessor's performance on real Māori words, which may contain such additional cues to morphological structure, to its performance on artificially constructed pseudo-Māori words, which are governed by the same patterns of statistical recurrence of morphs but lack any additional cues to morphological structure. This analysis reveals the extent to which such additional cues exist in real Māori and the extent to which they present issues for Morfessor. In doing so, it generalizes conclusions from Section 3 that the suitability of Morfessor to a particular language – and, by extension, the extent to which statistical learning by non-speakers of that language may be based purely on tracking of statistical recurrence – is dependent upon the morphological structure of the language.

4.1 Data

In this analysis, we focus entirely on words that follow Morfessor's underlying assumption of concatenativity. We do not include words that invoke templates at the word or morph level, since the analysis

in Section 3.2 already established that Morfessor underperforms in the presence of such templates.

The analysis is based on the 'polymoraic' group of Panther et al. (2024), excluding words with morphs containing more than 3 syllables. This includes a total of 1,292 words, comprising 1,199 of the 1,204 compounds that we analyzed in Section 3, as well as an additional 93 words that Oh et al.'s (2020) MS analyzed as simplex.

For the analysis of pseudo-Māori, we generated 1,000 different sets of 1,292 words each through concatenating morphs, based on the statistical properties of the 1,292 real Māori words (see Section 4.2). For each set, the generative process provided us with ground-truth segmentations, which we compare to those provided by a Morfessor model trained over the set. For the analysis of real Māori, we similarly compare the gold standard MS segmentations of the 1,292 words to those provided by a Morfessor model trained on those words (as opposed to the full lexicon from Section 3.1).

4.2 Methods

To generate each set of pseudo-Māori words, we used the same probabilistic process as is assumed by Morfessor's underlying generative model. This process works in a bottom-up fashion across several structural levels, first concatenating phonemes into syllables, then concatenating syllables into morphs, and finally concatenating morphs into words. Types of one level are drawn with replacement from an inventory, according to an inverse power law (Zipfian) probability distribution, and concatenated to form a type of the next level. The types at each level are unique: if a proposed type already exists, a new one is generated instead.

We generated each set of pseudo-Māori words with constraints based on real Māori, in two main ways. First, we constrained the pseudo-Māori words to have the same statistical recurrence properties as real Māori, by using an inventory probability distribution at each level that was inferred from the set of real Māori words (see Appendix A for details). Second, we constrained the types at each level to have the same form properties as real Māori. Specifically: at the phoneme level, we used the same 10 consonants and 10 vowels as real Māori (Section 2.3); at the syllable level, we only generated syllables of shape CV and V; at the morph level, we generated the same number of monosyllabic, disyllabic, and trisyllabic morph types (respectively) as there are in the real Māori set of words; and at the word level, we used each real Māori word as a template for a pseudo-Māori word, ensuring that they matched in terms of the number of morphs and the number of syllables in correspondingly-ordered morphs.

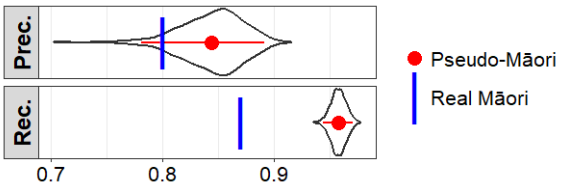

Figure 1: Distributions of macro-averaged precision and recall for Morfessor's segmentations of 1000 sets of pseudo-Māori words, in comparison to its performance on corresponding words from real Māori (blue lines). Red points show mean performance on pseudo-Māori and red lines show 95% percentile intervals.

As in Section 3.2.1, our analysis is based on comparing Morfessor segmentations to a gold standard. We again use macro-averaged boundary precision and recall as the metric for this comparison.

4.3 Results

Figure 1 shows the distributions of macro-averaged precision and recall for Morfessor's segmentations on the 1000 sets of pseudo-Māori words, together with the precision and recall for its segmentations of the corresponding real Māori words (when training is restricted just to those words). It is immediately apparent that recall is higher than precision, indicating occurrences of oversegmentation that are not balanced by undersegmentation as was the case in Section 3.2.2. This is likely a consequence of the training set being much smaller (1,292 words as opposed to 19,595); since the same pattern is seen across real Māori and psuedo-Māori, it does not appear to reflect influences of non-statistical cues to morphological structure.

Morfessor is better able to accurately segment pseudo-Māori than real Māori. Numerically, both precision and recall are higher for pseudo-Māori (mean precision: 0.84; recall: 0.96) than for real Māori (precision: 0.80; recall: 0.87). The advantage for pseudo-Māori is especially strong for recall, where performance on all 1,000 sets of words far exceeds that on real Māori. This strong advantage in recall is not driven by increased oversegmentation of pseudo-Māori relative to real Māori, because it is not accompanied by a concomitant disadvantage in precision; rather, it reflects the fact

that boundaries in pseudo-Māori are cued by recurrence statistics, which Morfessor tracks. That is, Morfessor is best able to segment words when they come from a language that closely adheres to the statistical principles of structure that it assumes.

It follows that Morfessor's worse performance on real Māori is likely due to failure to identify boundaries that are cued by something other than morph recurrence statistics. This result therefore confirms the suggestion from Section 3.2.2 that the morphological structure of Māori may have alternative cues, though it does not indicate precisely what they may be. Past research has shown that NMS are sensitive to cues such as bimoraic templates and the presence of long vowels in the segmentation of compounds (Panther et al., 2024), and it is likely that this sensitivity explains why they were more successful at segmenting compounds than affixed words in Analysis 1A.

5 Discussion & conclusions

We have examined morphological segmentations of Māori by Morfessor and non-Māori-speaking New Zealanders (NMS), across words formed through a variety of morphological processes, to assess the ways in which they are affected by structural factors and the extent to which they have such effects in common. Our results show that both learners are affected by linguistic structure. In some circumstances, they are affected similarly; for example, both are successful in segmenting words formed by concatenative morphological processes (Analysis 1A), especially when highly frequent morphs are involved (Analysis 1B). In other circumstances, they are affected in opposite ways; for example, Morfessor suffers decreased segmentation performance on words that are formed via templatic processes (Analysis 1A) or that cue morphological structure by means other that statistical recurrence of forms (Analysis 2), whereas NMS see increased performance in such cases.

These similarities and differences are important when considering the nature of human statistical learning of morphological segmentation. Since Morfessor's learning is underpinned by a set of well defined assumptions and principles (Section 2.2), the extent to which its performance aligns with that of NMS may be taken to reflect the extent to which NMS' learning is underpinned by those same assumptions and principles. The similarities affirm that NMS undergo statistical learning, identifying and extracting statistically recurrent forms to build a memory-store of morphs. At the same time, the differences show that learning for NMS does not just involve tracking statistical recurrence, but also involves inducing abstract templates about the formation of words and the shapes of (allo)morphs, as well as developing sensitivities to prominent features such as the presence of long vowels (Panther et al., 2024). These findings echo results showing that adults and infants attend to phonological templates when learning to segment artificial languages through incidental exposure (Peña et al., 2002; Marchetto and Bonatti, 2013).

On a practical front, the similarities and differences in the segmentation performances of Morfessor and NMS suggest that human statistical learning of morphological structure can be appropriately modeled by unsupervised machine learning, but perhaps only to a first approximation, depending on the underlying assumptions of the model. When the morphological structures closely follow those assumed by the model, the morphs that the model learns can reflect the cognitive units that humans seem to operate over (e.g., Virpioja et al., 2018; Lehtonen et al., 2019). But when morphological structures vary too widely from those assumed by the model – either within a language, based on words formed by different processes, or across languages – there is the potential for the model to miss factors that are salient to humans but that it is not equipped to handle. This is especially important as different models have different underlying assumptions, which can respond differently to variation in morphological structure (Loukatou et al., 2022).

The differences in the segmentation performances of Morfessor and NMS across words of different morphological structures not only inform the use of unsupervised morphological segmentation models as cognitive models, but also highlight potential factors that could be incorporated into segmentation models to improve their results. For example, inspired by the observation that reduplication templates are salient to humans but not to Morfessor, Todd et al. (2022) show that adding reduplication templates to Morfessor improves its ability to find reduplication in Māori words. Similarly, future research that dissects NMS' underlying learning mechanisms could reveal additional generalizable factors that help improve the cross-linguistic applicability of unsupervised models.

Limitations

While we believe our results to be informative about the effect of language structure on the construction of the NMS proto-lexicon, there are several limitations that could be addressed in future work to clarify and extend them.

First, the gold standard data may not strictly reflect morphological segmentations. One reason for this is that the word-segmentation task through it was obtained taps a form a meta-linguistic knowledge that may not be directly accessible in a consistent manner. However, we do not think this to be a major concern, given that past work using the same task in English (Needle and Pierrehumbert, 2018) found that participants' segmentations matched the underlying morphological structure 88% of the time, and given that we filtered the words used in the analysis to only include those where the gold-standard segmentation is consistent with the assumed morphological structure. We also do not see a better option than eliciting meta-linguistic judgments in this case: the largest group of morphologically complex words in Māori is compounds (Bauer, 1993; Todd et al., 2019), which are not decomposed in any dictionary or large word list of which we are aware.

Second, and relatedly, the gold standard data may contain idiosyncracies, since it was provided by a single MS. While the MS was instructed to split words into parts in a way that they think most Māori speakers would agree with, it is extremely unlikely that their segmentations would all be universally shared. To address this limitation, it would be necessary to repeat the word-segmentation task with many more MS, like we did for NMS.

Third, our comparison of Morfessor and NMS may be complicated by differences between them. For example, Morfessor has perfect memory about all forms of the language and its segmentations of them, but NMS are unlikely to have encountered all words of the language, let alone remember those encounters. Similarly, Morfessor is trained on isolated unique types, whereas NMS experience connected tokens. Morfessor's knowledge is also limited to its Māori training data, whereas NMS also have knowledge of at least one other language (English). It remains to be seen how well Morfessor does when trained on data that resembles what NMS are exposed to, including connected tokens of both Māori and English, and how it may be affected by memory constraints.

Acknowledgements

We thank Forrest Panther for providing the NMS segmentation data. This work used computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California Nano Systems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara.

References

Richard N. Aslin. 2017. Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1–2):e1373.

Laurie Bauer and Winifred Bauer. 2012. The inflection-derivation divide in Māori and its implications. *Te Reo*, 55:3–24.

Winifred Bauer. 1993. *Maori*. Routledge, London.

Mary Teresa Boyce. 2006. *A Corpus of Modern Spoken Māori*. Unpublished doctoral dissertation, Victoria University of Wellington.

Hermann Bulf, Scott P. Johnson, and Eloisa Valenza. 2011. Visual statistical learning in the newborn infant. *Cognition*, 121(1):127–132.

Mathias Creutz 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.

Paul de Lacy. 2003. Maximal words and the maori passive. In *Proceedings of AFLA VIII: The eighth meeting of the Austronesian formal linguistics association*, volume 44, pages 20–39. MIT Linguistics Dept.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 900–910.

Michael C. Frank, Joshua B. Tenenbaum, and Edward Gibson. 2013. Learning and long-term retention of large-scale artificial languages. *PLOS ONE*, 8(1):e52500.

William A. Gale and Geoffrey Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

Judit Gervain, 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):14222–14227.

Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-decker, Gilles Adda, Hélène Maynard, Annie Rialland, and Inria Grenoble. 2018. Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In *Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–42.

Ray Harlow. 2007. *Māori: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Elizabeth K. Johnson. 2016. Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics*, 2(1):391–412.

Mark Johnson and Thomas L. Griffiths. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*, pages 641–648.

Peter J. Keegan. 1996. *Reduplication in Maori*. Unpublished MA thesis, University of Waikato.

Jeanette King, Margaret Maclagan, Ray Harlow, Peter Keegan, and Catherine Watson. 2011. The MAONZE project: Changing uses of an indigenous language database. *Corpus Linguistics and Linguistic Theory*, 7(1):37–57.

Victor Krupa. 1968. *The Maori Language*. Nauka, Moscow.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. arXiv preprint arXiv:1804.10959.

Minna Lehtonen, Matti Varjokallio, Henna Kivikari, Annika Hultén, Sami Virpioja, Tero Hakala, Mikko Kurimo, Krista Lagus, and Riitta Salmelin. 2019. Statistical models of morphology predict eye-tracking measures during visual word recognition. *Memory & Cognition*, 47:1245–1269.

Georgia Loukatou, Sabine Stoll, Damian Blasi, and Alejandrina Cristia. 2022. Does morphological complexity affect word segmentation? Evidence from computational modeling. *Cognition*, 220:104960.

Erika Marchetto and Luca L. Bonatti. 2013. Words and possible words in early language acquisition. *Cognitive Psychology*, 67(3):130–150.

John C. Moorfield. 2011. *Te Aka: Māori-English, English-Māori Dictionary*, 3rd edition. Pearson, Auckland.

Jeremy Needle and Janet B. Pierrehumbert. 2018. Gendered associations of English morphology. *Laboratory Phonology*, 9(1):14.

Céline Ngon, 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.

Yoon Mi Oh, Simon Todd, Clay Beckner, Jennifer Hay, and Jeanette King. 2020. Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific Reports*, 10(1):22318.

Yoon Mi Oh, Simon Todd, Clay Beckner, Jennifer Hay, and Jeanette King. 2023. Assessing the size of non-Māori-speakers' active Māori lexicon. *PLoS ONE*, 18(8):e0289669.

Forrest Panther, Wakayo Mattingley, Jennifer Hay, Simon Todd, Jeanette King, and Peter J. Keegan. 2024. Morphological segmentations of non-Māori speaking New Zealanders match proficient speakers. *Bilingualism: Language and Cognition*, 27(1):1–15.

Forrest Panther, 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).

ʻŌiwi Parker Jones. 2008. Phonotactic probability and the māori passive. In *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 39–48.

Marcela Peña, Luca L. Bonatti, Marina Nespor, and Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science*, 298(5593):604–607.

Bruna Pelucchi, Jessica F. Hay, and Jenny R. Saffran. 2009. Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80(3):674–685.

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

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Aku Rouhe, Stig-Arne Grönroos, Sami Virpioja, Mathias Creutz, and Mikko Kurimo. 2022. Morfessor-enriched features and multilingual training for canonical morphological segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 144–151.

Jenny R. Saffran. 2001. Words in a sea of sounds: the output of infant statistical learning. *Cognition*, 81(2):149–169.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

Tuomas Teinonen, 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.

Simon Todd, 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.

Simon Todd, 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*, pages 12–22.

Simon Todd, Jeremy Needle, Jeanette King, and Jennifer Hay. 2019. Quantitative insights into Māori word structure. Paper presented at the Annual Meeting of the Linguistic Society of New Zealand.

Sami Virpioja, Minna Lehtonen, Annika Hultén, Henna Kivikari, Riitta Salmelin, and Krista Lagus. 2018. Using statistical models of morphology in the search for optimal units of representation in the human mental lexicon. *Cognitive Science*, 42(3):939–973.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical report, Department of Signal Processing and Acoustics, Aalto University, Helsinki.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

John N. Williams. 2020. The neuroscience of implicit learning. *Language Learning*, 70:255–307.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

A Generating pseudo-Māori: Details

This appendix describes the process through which we inferred statistical recurrence properties of Māori, to use in the generation of pseudo-Māori.

We derived inventories at each level – unique phonemes, syllables, morphs, and words – from the segmentations provided by Oh et al.'s (2020) fluent Māori speaker. To get the frequency distribution over types at one level, we counted occurrences within unique types at the next level. That is, we counted the number of unique syllables that each phoneme occurred in; the number of unique morphs that each syllable occurred in; and the number of unique words that each morph occurred in. We sorted each distribution by count, to obtain rank and frequency for each type, and fit an inverse power law $f(x) = ab^{-x}$ to predict frequency from rank, using nonlinear least squares.

To sample in the generative process, we sorted the types in random order and treated those orders as ranks, overlaying the frequency from the inverse power law and then normalizing to obtain a probability distribution.