A Quest for Paradigm Coverage: The Story of Nen

Saliha Muradoğlu[♣♠] **Hanna Suominen**[♣♢] **Nicholas Evans**[♣♠]
[♣]The Australian National University (ANU) [♢]University of Turku
[♠]ARC Centre of Excellence for the Dynamics of Language (CoEDL)
Firstname.Lastname@anu.edu.au

Abstract

Language documentation aims to collect a representative corpus of the language. Nevertheless, the question of how to quantify the comprehensive of the collection persists. We propose leveraging computational modelling to provide a supplementary metric to address this question in a low-resource language setting. We apply our proposed methods to the Papuan language Nen. Nen is actively in the process of being described and documented. Given the enormity of the task of language documentation, we focus on one subdomain, namely Nen verbal morphology. This study examines four verb types: copula, positional, middle, and transitive. We propose model-based paradigm generation for each verb type as a new way to measure completeness, where accuracy is analogous to the coverage of the paradigm. We contrast the paradigm attestation within the corpus (constructed from fieldwork data) and the accuracy of the paradigm generated by Transformer models trained for inflection. This analysis is extended by extrapolating from the learning curve established to provide predictions for the quantity of data required to generate a complete paradigm correctly. We also explore the correlation between high-frequency morphosyntactic features and model accuracy. We see a positive correlation between high-frequency feature combinations and model accuracy, but this is only sometimes the case. We also see high accuracy for low-frequency morphosyntactic features. Our results show that model coverage is significantly higher for the middle and transitive verbs but not the positional verb. This is an interesting finding, as the positional verb paradigm is the smallest of the four.

1 Introduction

A key question in studying language is: when do we have enough data to fully understand the system? This is especially important in language documentation. As Himmelmann (1998) states, *'the aim of language documentation is to provide a comprehensive record of the linguistic practices characteristic of a given speech community.'*. Bird (2015) extends this by asking, *'If a comprehensive record is unattainable in principle, is there a consensus on what an adequate record looks like. How would you quantify it?'*.

Honouring their formulation, Baird et al. (2022) label this the 'Himmelman-Bird' problem.[1] In their paper, the authors strive to explore this Himmelman-Bird problem for the inventory of phonemes, which are the subdomain of language with the smallest and hence most frequently-occurring units. They set the bar even lower by simply requiring that at least one allophone of each phoneme occur. They then examine how much text it might take to capture a language's entire phoneme inventory, drawing on a sample of 137 distinct languages, some with additional dialectal or register variety taking the total to 158 speech varieties. Full 'coverage' is achieved, for a given domain of language (say, its phoneme inventory) and a given corpus, it there is at least one incidence of each relevant unit (in this case, each phoneme) in that corpus.

Here we strive to follow a similar route for morphemes and their respective allomorphs, while still posing the problem in its simplest and hence most easily-satisfied form: we look just at verbs, and we restrict ourselves to one representative lexeme (the commonest) in each of the four main morphological classes – see below.

The goal of collecting a representative sample has permeated many fields, from biology to sociology. Researchers have explored the idea of having a gold standard process for collecting all required components to describe a system. For example, if we wanted to gather all the phonemes for English, the 'Rainbow Passage' by Fairbanks (1960) may be chosen. The first four lines of the passage cap-

[1]This is akin to the problem of corpus representativity.

Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 74–85

ture all phonemes for English. In morphology, we can discuss the idea of collecting all principal parts (Finkel and Stump, 2007) to construct the entire paradigm.

This idea presents as a great solution to the difficulty faced by low-resource languages and, more specifically, language documentation. However, one caveat is the system knowledge required for designing such a task. For example, how might a linguist know all the phonemes before beginning their in-field analysis and recordings? Accordingly, we make the distinction between heuristic and attestation coverage.

The first refers to the discovery stage of a language, leading to a sketching of the dimensions of its design space - the logical space of all its possibilities in a particular domain, such as verbal inflections – through discovering the dimensions where it encodes contrasts (say 'dual number', 'future imperative', 'imperfect aspect'), and mapping out the ways these interact (say 'future imperfective dual imperative', as in Nen *nandowabe* 'you two should be talking later on!' (Evans, 2019). The latter describes the scenario where a description exists, and the aim is to collect examples of language within the denoted design space.

The concept of a 'whole language' is so vast and heterogenous that it is not operationally useful for many linguistic or practical purposes. To explore this question, we consider a particular component of language, inflectional morphology on the verb. We base our study on modelling morphological inflection in the Nen language and examine the attestation coverage observed in the transcribed natural spoken corpus and inflection models built on the same data.

In this paper, we address the following questions: (1) How can we test the degree to which a linguistic subsystem exhibits coverage in a given corpus (2) How does the model coverage compare with the corpus? (3) Does corpus frequency relate to model accuracy? (4) Can we use model-based learning curves to predict the data required for complete coverage?

We propose a test case for the model that asks to predict a complete paradigm, i.e. the complete multidimensional array of inflected forms – English is too morphologically impoverished to furnish a good example (the best is with the copula to be: {*am, (art), is, are; was, were; (to) be; being*}. Our results indicate that the generalisations afforded by the Transformer model yield better coverage than the natural corpus. Furthermore, we explore two separate correlations of the high dimensional axes of Nen verbs; the undergoer and agent combinations and the agent and Tense, Aspect, and Mood (TAM) combinations. While frequent features tend to be captured correctly by the model, surprisingly, so are some low-frequency forms. Finally, we use learning curves to predict the data needed for 100% coverage.

2 Related Work

To our knowledge, only two prior computational studies of Nen exist. Muradoglu et al. (2020) presents a finite-state description, while (Muradoğlu et al., 2020) explores the use of neural architecture, to model Nen verbal morphology. The latter is based on two high performing submissions in the SIGMORPHON–CoNLL 2017 Shared Task (Cotterell et al., 2017). Between the two approaches, the finite-state description achieves a higher accuracy across the corpus. However, we note that the accuracies reported are not directly comparable given the ongoing development of the corpus.

Despite the performance difference, we opt to use a neural approach to enlist the aid of its generalising ability. Moreover, the statistical nature of these models make the intersect with corpus linguistics an object of interest. Specifically, we use a Transformer (Vaswani et al., 2017) based model. Transformers have been successful in capturing complexities of phonological and morphological details (Pimentel et al., 2021; Kodner et al., 2022), often achieving state-of-the-art performance. Over the years, the inflection task has been extended to many languages, including other complex morphological systems such as Murrinh-Patha, Kunwinjku and Seneca.

3 The Nen Language

Nen is a Papuan language of the Morehead-Maro (or Yam) family (Evans, 2017). It is spoken as a native language in the village of Bimadbn in the Western Province of Papua New Guinea (Evans, 2015, 2019). Most Nen speakers are multilingual, typically speaking several of the neighbouring languages.

Verbs in Nen are notoriously complicated and are described as the most complicated word-class in Nen (Evans, 2015, 2019). They can be grouped

in several ways, either as prefixing and ambifixing or by further breaking down the inflection patterns. Prefixing verbs consist of the copula (and its derivatives 'go'/'come'/'have'), 'to walk' and positional verbs. Another distinguishing feature of prefixing verbs, is the lack of infinitives. Both ambifixing and middle verbs form infinitives through suffixing *-s* to the verb stem. In this study, we have listed the prefixing verb lemmas as the verb stem. Ambifixing verbs can be separated into middle and transitive verbs. Here, we separate the verb types beyond the prefixing and ambifixing categories as the corresponding paradigms are distinct. We provide details for the verbs we track below.

3.1 Copula

The copula is a special case for our test, in that we test the generation of a partial paradigm as the model would have seen several forms of the copula. We note that this verb, together with its directional counterparts 'come' and 'go'. The come/go paradigms are built using the copula with the addition of directional prefixes, is the most frequent verb type in the corpus. The copula paradigm consists of 40 unique forms. See Evans (2014) for full paradigm.

3.2 Positional

Verbs in the positional class fall into two main types: posture and position proper (Evans, 2015). For example, *mängr* 'be lying in a jumble' and *érningr* 'be in hiding' or spatial position in relation to some frame of reference like *pingr* 'to be high (typically inanimate)'. So far, 45 verbs have been recorded. Verbs of this class have special stative suffixes *-ngr* for non-dual and *-aran* (dual). They exhibit properties of prefixing verbs: they do not have infinitives and cannot form present imperative (Evans, 2014).

3.3 Middle

Middle and transitive verbs have the same TAM paradigm. Aside from valency, the distinction between the two is that the middle verbs have a dummy prefix with no semantic meaning other than to note that they are middle verbs. This prefix does not mark an argument like other verb types. In rare cases, middle verbs use the undergoer prefix slot to index large plurals. Example verbs of this type include *owabs* 'to speak' or *angs̄* 'to return'. Both these verbs are ambifixing, but the prefixal slot is restricted to {*n-*} (α–series), {*k-*} (β–series), {*g-*} (γ–series).

3.4 Transitive

By contrast, transitive verbs utilize both prefixes and suffixes to mark person and number. Examples of this verb type include *yis* 'to plant' and *waprs* 'to do' These verbs allow for full prefixing and suffixing possibilities. The prefix set is divided through the use of the same arbitrarily labels α, β, and γ, as the middle verbs. Instead of the middle verb marker, transitive verbs allow for person/number undergoer marking. These dummy indices do not carry specific semantic values until they are unified with other TAM markings on the verb.

Evans (2016) provides the canonical paradigms for the undergoer prefixes, thematics and desinences. Suffixes are constructed by combining the corresponding thematic and the desinence. The future imperative construction is a special case, where an additional future imperative prefix is required (Evans, 2015).

3.5 Directional

Following the undergoer prefixes, a directional prefix slot is available. This can be filled with {*-n-*} 'towards', {*-ng-*} 'away' or left empty to convey a directionally neutral semantic.

Consider the copula verb *m* 'to be', when marked for direction the resultant forms are as follows: *y-n-m* '(s)he coming (towards speaker)', *y-ng-m* '(s)he is going (away from speaker)'. Note the speaker centric frame of reference.

4 Data

The Nen corpus is made of 44 individual texts that were naturalistically recorded in the field. This amalgamates to approximately 8 hours of spoken text or over 30,000 words. This is filtered to over 6,000 verb instances representing 2,282 forms. Some of these forms are the same, with different feature combinations due to syncretism or polysemy. For example, the sequence *yn-* can be parsed in two ways. It can either mean the prefix *yn-* coding first person nonsingular undergoer for the α series or *y-n* the third singular undergoer with the ventive (towards) directional. Each of these instances are treated separately to expose the model to all possible meanings.

A large portion of the texts in the corpus are

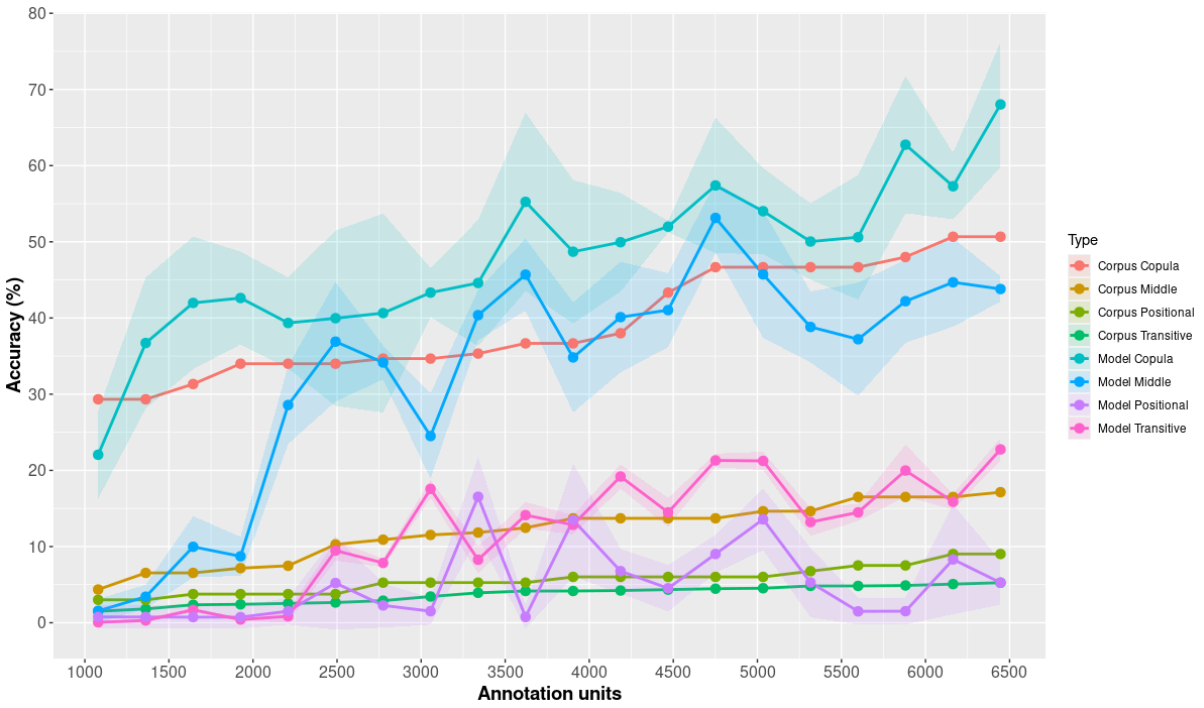

Figure 1: The coverage growth for four verb types in Nen, reported as a function of Annotation units (within corpus), where 'annotation units' are audibly-demarcated units in the flow of speech (typically by pause breaks). In our corpus, on average there is one verb per annotation unit, making annotation units a reasonable proxy of how often we would expect verbs to occur. The corpus accounts follow *akingr* 'to be standing' for the positional, *owabs* 'to speak' for the middle and *räms* 'to do/give' for the transitive. The confidence bands reported on the model results are calculated based on a 4-partition variance. The full Nen corpus currently consists of 6,446 annotation units. The starting point is 1,079 as this roughly corresponds to 382 (100 train + 282 dev) instances.

coconut interviews[2], these typically involve so-called biographical questions (parent names, place of birth etc), and questions about coconut trees that belong to the interviewee. This type of text was chosen as it can include a variety of tense - whether someone has planted or will plant a coconut tree - and is a topic that easily inspires conversation from locals. Although, these do not constitute a genre in the traditional sense, they do exhibit characteristic features, such as a high token count of the verb *yis* 'to plant' and third person non-past copula *ym*. The remaining texts range from anecdotal stories, folk tales, other narratives or procedural explanations.

5 Experiment

We contrast the corpus-based account of the Nen verbal paradigm to that modelled by a Transformer model (Wu et al., 2021). Our study is conducted in two parts: first, we follow the attestation coverage of the paradigm for one representative verb for each type in the corpus. Second, we train Transformer models to generate a complete paradigm for an unseen (barring the copula) verb for each type with incremental amounts of data. We establish a learning/coverage curve for each method (Anzanello and Fogliatto, 2011; Viering and Loog, 2022). We use the term coverage here to mean the percentage of cells observed in the corpus or correctly predicted by the models out of the entire language design space.

5.1 Corpus-based Account

Here we present a corpus account of paradigm coverage. For each of our four verb types, we follow the trajectory of the lexeme.[3] As it happens the top three verbs, by frequency, are the copula (most frequent at 80.46 IPT (Items per thousand)[4], the middle verb *owabs* 'to speak' (Second most frequent lexeme in the corpus, 6.83 IPT) and the transitive

[2]See Evans (2020) for more details.

[3]Where a lexeme is a'dictionary word', i.e. the citation form of a word used in a dictionary, and uniting all its inflected forms. Thus the lexeme *run* unites the inflected forms *run, runs, ran* and *running*. In Nen the number of inflected forms per lexeme is much larger, as we shall see below.

[4]The more common metric is IPM (items per million) but given that the size of the Nen corpus is in order of thousands, we report these figures in IPT.

verb *räms*'to do/give' (Third most frequent lexeme in corpus, 6.46 IPT). We then have to descend some way down the frequency list before reaching our highest-frequency positional verb, namely *akingr* 'to be standing' (16th most frequent lexeme, 1.83 IPT).

For our four verbs, we then collate all distinct forms of the verb in question, tracking for where in the corpus it is encountered. For example, for the verb *akingr*, the first form *yakingr* is encountered at the 223rd annotation unit, the second *ynakiaran* at 242nd and so on. The texts within the corpus are concatenated, and the same order of the text is preserved for each analysis.

The copula verb *m* is included in both training and test since it makes up for a large portion of the existing corpus and occupies the top 5 most frequent forms. It is the most frequent lexeme (80.46 IPT). This scenario can be seen as a more straightforward case, as 62.5% of the copula paradigm (without the directional prefix) is attested in the complete 2,000 instance training data. So the model needs to reproduce these forms with the directional prefixes. The remaining three verb types are not encountered in training time, barring the stem.

5.2 Model-based Account

We train models like an 'inflection' task in the SIGMORPHON shared tasks (Kodner et al., 2022), with tags identifying morpho-syntactic categories. The system is asked to produce the inflected form given the lemma and morpho-syntactic tags. For example, ⟨`owabs, V;IPFV.NPHD;1SGA;M;`α`, nowabtan`⟩ or the English equivalent ⟨`talk, V;V.PTCP;PRS`[5]`, talking`⟩.

We additionally account for the copy bias reported in (Liu and Hulden, 2022) by including the three[6] (see Section 5.2.2 for details) lemmas considered during test time in the training set.

Each model is trained using a character-level Transformer (Wu et al., 2021). This model has been used as the neural baseline for the SIGMORPHON shared task on morphological inflection[7].

We train models based on a Zipfian sampling strategy, as corpora obey Zipf's law at all sample sizes (Baayen, 2001; Blevins et al., 2017). The dev set is determined as the least frequent 282 forms and is kept the same for every experiment. The distribution is calculated from the existing corpus study (Muradoğlu, 2017). We train at 100 training sample intervals, ranging from 100 to 2,000 instances.

Prior work has explored the difference between random and Zipfian sampling. For example, Muradoğlu et al. (2020) examined the difference and reported that random selection yielded better results (or a faster coverage rate). However, given our research question, what random sampling means for language documentation is unclear. With many of the corpora built by field linguists built upon a combination of standard field method practices and anthropological story gathering, the type of data collected is hardly random. As such, the model results presented in this paper are based on Zipfian sampling.

5.2.1 Design of Test

We propose a modified test case to measure paradigm coverage of the model. A lexeme is chosen for each verb type and tested for each cell or unique morphosyntactic description (MSD).

The choice of lexeme is motivated by how regular the inflection of its particular phonotactics are. With the purpose of testing generalisability, it follows that our case study verbs are regular. Although we note that limitations of this approach, namely the variation of morphs across certain phonological properties of the stem (e.g., vowel harmony).

Given resource and access limitations we have utilised the finite-state grammar for Nen (Muradoglu et al., 2020) to generate full paradigms for the positional and transitive verbs, these paradigms are later examined by a language expert. The middle verb test is based on a full paradigm that was previously verified with Nen speakers. The full copula paradigm and its directional variants are sourced from the forthcoming grammar of Nen.

In a sense our suggested test for coverage is similar to the wug test in the SIGMORPHON shared tasks (Kodner et al., 2022), but rather than general production processes of nonce words we are interested in generating complete paradigms.

5.2.2 Meet the Verbs

***m* 'to be'** The copula paradigm consists of 40 unique forms. The come/go paradigms are built using the copula with the addition of directional prefixes.

[5]Present participle

[6]Since the model is already exposed to the copula during training time, it does not need to be included again.

[7]Model parameters follow (Wu et al., 2021).

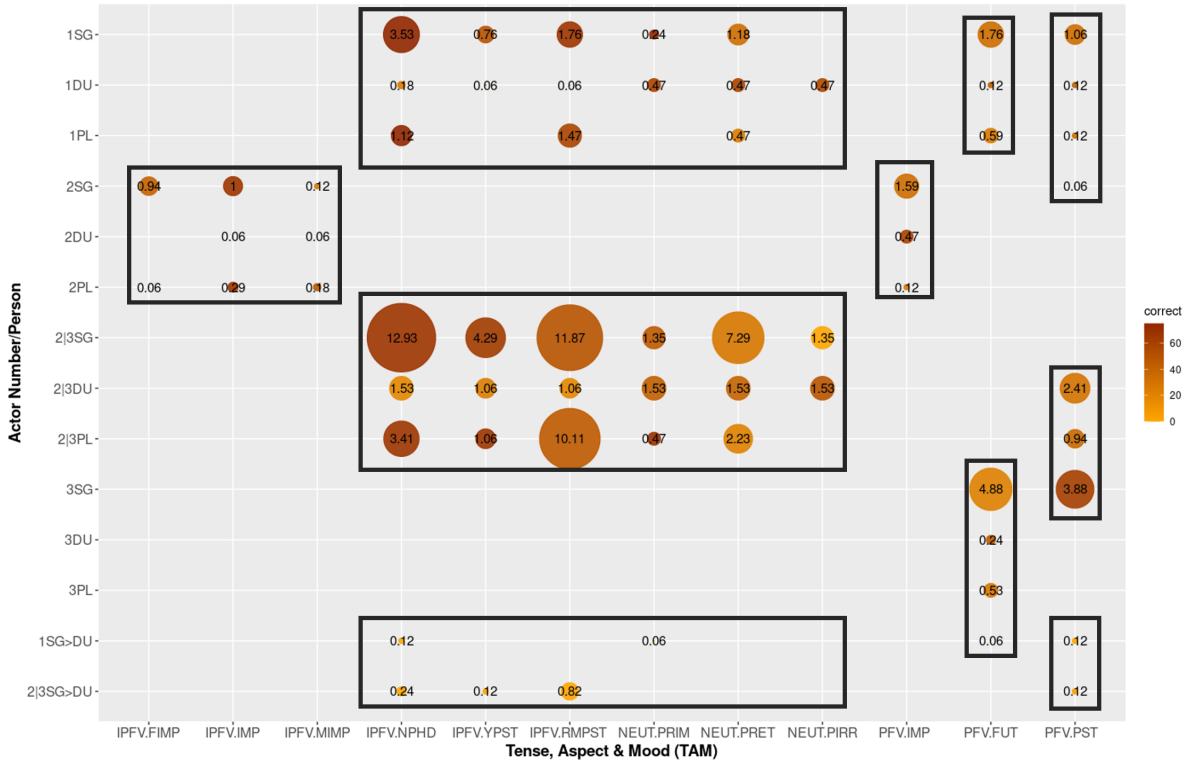

Figure 2: Bubble plot showcasing the frequency correlation between TAM and agent person number, reported numbers are percentage of corpus with the TAM/agent features. Navy lines indicate available cells described by the language design space. Note that the second and third persons are typically display syncretism except in the perfective past. See appendix A for details on TAM categories. The darker the colour (towards a blood orange) the more proficiency the model displays. Conversely the lighter the colour (orange) the more the model struggles to produce a correct form with the corresponding features.

***pingr (n-du)/piaran (du)* 'to be high/elevated'** Depending on the vowel of the stem ('i' in this case), the 2|3nsg prefix is e-, e.g., *epingr* 'you two/they two are up high'.

***armbs* 'to climb'** As with all middle verbs, *armbs* begins with a vowel. It is somewhat similar to the most common middle verb in the corpus *owabs* 'to speak', with a shared **b** before the infinitive marker *-s*. In addition to exhibiting regular inflection, the forms have been verified by native Nen speakers.

***wambaes* 'to sniff'** There are a few key points to note for this verb. When verb infinitives end with a dipthong (e.g. ae) before the final s, the dipthong is shortened in the non-dual (e.g., *wakaes* 'to look at' but *yakatan* 'I look at him/her'), but in the dual the full diphthong is present and also a dual-marking -w- which only occurs in such environments, e.g., *yawakataewn* 'I look at the two of them', *yakataewm* 'we two look at him/her'.

The most notable verb that is similar in phonological structure is *wakaes* 'to see'. The corpus contains 36 unique forms for *wakaes*.

6 Results and Discussion

A full paradigm for one verb is unlikely to be encountered in natural speech, or language learning contexts (Chan, 2008; Blevins and Blevins, 2009). Although the focus of this paper is not language learning, the sparsity of paradigm coverage observed in these contexts is equally relevant here. Based on various well-known corpora, Chan (2008) shows that languages with larger verbal paradigms exhibit lower coverage. Most notably, the only language with full coverage of its verbal paradigm is English, which only has six verbal forms. By contrast, Finnish has 365 verb forms and only a 40.3% saturation even though the corpus size is almost double (2.1 million words compared to the Brown corpus of 1.2 million words) that of the English counterpart.

Muradoğlu (2017) reports on the bleak data requirements to record each cell of the transitive verb in Nen. Here we have utilised the power of transformer models to leverage abstraction and statistical learning. Figure 1 shows that the model based

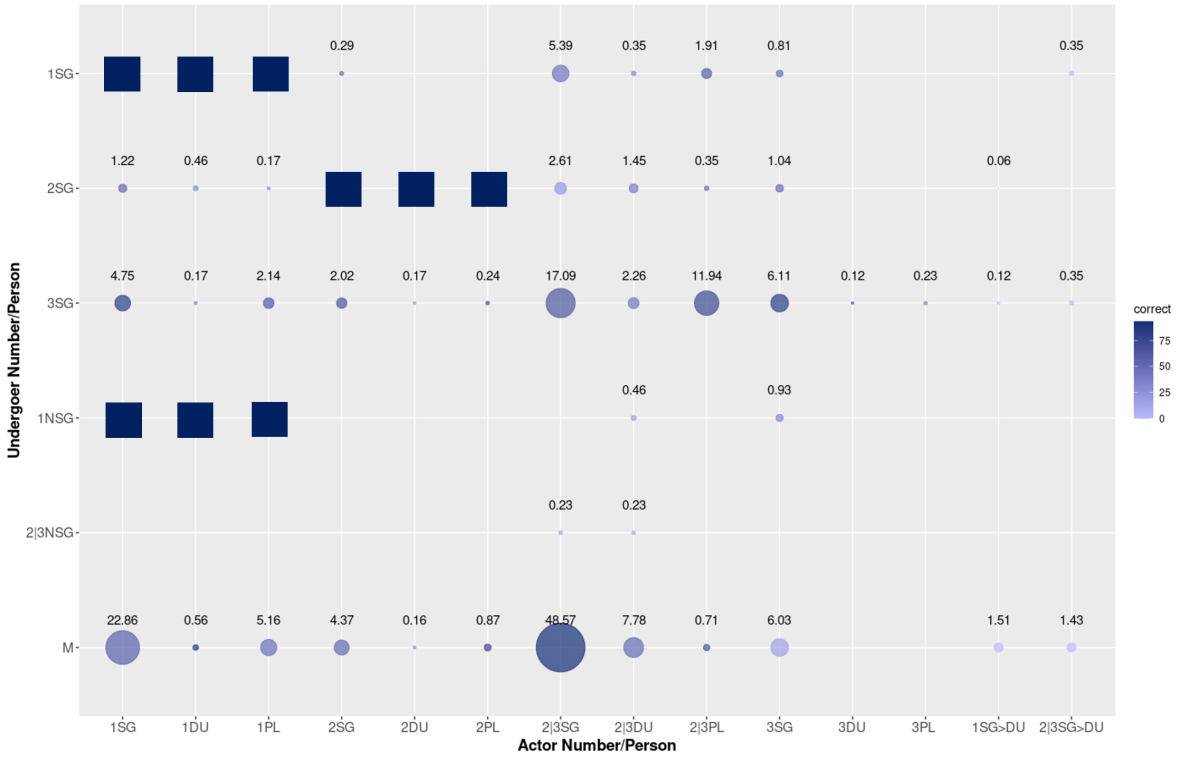

Figure 3: Relativised bubble plot of Actor and Undergoer person number for Nen. The navy blue blocks note the semantically disallowed combinations or in the case of first person acting on first person this meaning is achieved through reflexive constructions. The darker the colour (towards a purple) the more accurate the model is. Conversely the lighter the colour (lavender) the more the model struggles to produce a correct form with the corresponding features.

on the corpus does significantly better in terms of coverage. This suggests that while each combination might not be present in the corpus, the relevant information is. This typically parallels a mechanism utilised by field linguists to bootstrap the mapping of a linguistic paradigm since going through a complete paradigm for one particular verb is implausible. Instead, the circumstantial context primes language informants to showcase verbs of different semantic domains. The field linguist typically obtains part of the paradigm (either through elicitation or by natural means) for each verb. These fragments likely allow for a reconstruction of the entire paradigm. Dimensional independence allows the linguist to fill out parts of the paradigm. This task has been described as the paradigm cell filling problem (PCFC) Ackerman et al. (2009); Silfverberg and Hulden (2018); Liu and Hulden (2020).

Figure 1 shows the paradigm coverage across the four verb types in question. We contrast model-based coverage with a corpus-based account. In both instances, we follow the trajectory of one representative verb. For the model, the four test verbs are detailed in the Section 5.2.2. The corpus coverage curve follows *akingr* 'to be standing' for the positional, *owabs* 'to speak' for the middle, and *räms* 'to do/give' for the transitive verb. The model and the corpus follow *m* 'to be' since the copula verb is one entity.

The most observable behaviour shown in Figure 1 is the fluctuation across models trained across different training sizes. Although, in general, the growth is positive, we see a significant difference across each step. One explanation might be the skew within the samples added. In other words, the added examples negatively influence the generalisations built by the model. Another might be the model sensitivity to initial training data and data order. To account for the statistical variation, we report confidence bands for each verb type by measuring the variation in accuracy by dividing the test case for each verb into four random partitions. The partitions are randomly sampled as the test file is constructed in paradigmatic order. If the partitioning is performed sequentially, we might

	Corpus		Model		
	Annotation units	# of words	Training size	Annotation units	# of words
All	–	–	198,000	560,000	2,610,000
Transitive	154,000	716,000	34,000	97,000	451,000
Middle	44,000	205,000	4,000	12,000	55,000
Positional	40,000	188,000	3,000	10,000	45,000
Copula	11,000	53,000	3,000	10,000	46,000

Table 1: Extrapolated values based on the learning curve for both corpus and model-based coverage. The corpus's training size has been omitted as it does not bear any particular meaning. The numbers presented are rounded to the nearest thousand.

observe bias in one part of the paradigm, yielding large error margins.

The model shows greater coverage for the transitive, middle and copula verb types than the corpus account. Interestingly, the growth curve shows that the model-based account for positional verbs does worse than the corpus account. This is because the learning curve for the positional verb fluctuates substantially. The best-performing model for positional verbs is obtained with only 900 training examples (or 3,339 annotation units) at 16.5% coverage compared with the corpus account of *akingr* at 9% across the whole corpus. Given that the paradigm of the positional verb is the smallest among the four, we would have expected coverage to be high. A possible explanation for this might be that there are few instances of positional verbs in the corpus (26 distinct forms across seven lexemes) and, thus, the training set. We also observe looping errors as described in Shcherbakov et al. (2020), particularly for training sets below 1,000 instances.

We describe the coverage growth relative to annotation units to capture the data requirements for paradigm representation fully. The texts are segmented into annotation units to retain some of the contextual information surrounding the verb in question. These units are typically one complete sentence and most commonly correspond to a segment in ELAN (Sloetjes and Wittenburg, 2008). On average, 4.7 words per intonation unit, one of which is usually a verb. With 6,446 annotation units across the corpus, on average, for every 2.88 units, there is a distinct form encountered.

The model paradigm coverage is contrasted with that from the Nen spoken corpus. We make a point to situate the required data size for training the model (i.e., train + dev) with units that relate to the corpus to help highlight the distillation process. Typically, the model training size is measured in the number of instances. However, when collating a data set for a specific natural language processing (NLP) task – such as morphological inflection, the corpus is filtered from total words (assuming transcription exists) and later further distilled to types from tokens.

To address our third question, we analyse the frequency of the verb features along the TAM/Actor and Actor/Undergoer dimensions. We expect a strong correlation between highly frequent features in the corpus and the model accuracy for that slot. Figures 2 and 3 show the frequency of feature bundles. In both figures, the size of the bubbles corresponds to the frequency of the two sets of features in question (TAM and Actor or Actor and Undergoer). The saturation of the bubble shows how successful the model is in capturing the particular feature combination. The darker the bubble, the more likely the model will produce the correct corresponding form. These results are based on the model training with the entire training set available (2,000 instances).

As expected, both figures show a correlation between the bubble size (corpus frequency) and saturation (model accuracy). Nevertheless, there are cases where the corpus frequency is low, but the model proves to be proficient in producing the correct form. One such example is the imperfective imperative (ipfv.imp), the second person plural actor (which requires a prefix of the α series and the *-tang* suffix) makes up for 0.29% of the training data, but the model produces the correct form more than 66% of the time. One explanation might be that the rule's complexity and the chosen test verbs do not trigger allomorphic variants.

We note the morphophonological element of inflecting. While we have tried to choose regular verbs, they still exhibit a phonological layer. It is hard to disentangle such effects. One possible

future direction would be to choose a list of verbs across the categories presented here which exhibit the full range of phonological phenomena observed in Nen. For example, verbs that might trigger vowel harmony and the consequent allomorphs.

We further our analysis by providing a predictive quantity of data needed to reach 100% accuracy. We utilise scipy-based (Virtanen et al., 2020) extrapolation by treating the resultant coverage curve as a learning curve. The predictions presented here are optimistic; to ensure that the predictions are based on monotonically increasing functions, we ensure that:

$$A(AU') > A(AU)$$

where A is the accuracy, AU is the annotation units and $AU' > AU$. Given the predictions' variability, the numbers are rounded to the nearest thousand. Table 1 shows that the amount of data needed for the model to reach full coverage is significantly less than a corpus-based account. In some cases, such as the transitive and middle verb, the estimated quantity is over four times less. We expect these paradigms to benefit the most from generalising as they typically display regular inflection. Additionally, the paradigm size for both is substantial.

It is tempting to draw parallels between language learning and the analysis presented here. However, we remind readers that we base our predictions on one representative verb and focus on attestation coverage rather than heuristic coverage. Furthermore, we note that heuristic coverage would require a vastly more significant quantity of data. In addition, the numbers here are for one verb only, and it does not extend to include all parts of speech.

7 Conclusion

We propose 'coverage' as a new way to measure the comprehensiveness of a corpus for morphological paradigms. Here we present this application to Nen verbal morphology. This methodology can be extended to include other parts of speech or languages.

Our results show that using deep learning approaches, more specifically the Transformer architecture (Gillioz et al., 2020; Lin et al., 2022) allows us to exploit the generalisable parts of a paradigm and thus grant us a higher coverage. The model-based account yielded higher attestation for three of the four verbs considered. In an ideal setting, each inflection feature for each word would be observed and recorded naturally. However, this is an impossible feat in real-life. Using statistics-based modelling like the Transformer model allows us to synthesise forms based on examples encountered in the training data. As a result, the existing corpus can account for more of the system than a simple count within the corpus would suggest.

We have explored the basis of the conventional wisdom of higher frequency yielding better model performance. While this holds, we observe a positive correlation between high-frequency feature combinations and model accuracy; we also see that the model can correctly generate less frequent feature combinations as well.

We provide data quantity estimations based on the learning curves generated. These predictions are meant only as a guide rather than anything definitive, as they present an optimistic case defined by the enforcement of monotonicity.

The extension of our proposed methodology to other languages with diverse morphological characteristics remains an open direction for future work.

Limitations

One major limitation of the study presented here is the microscopic tracking of one representative verb. As mentioned earlier, one potential solution is to track several verbs of each inflection type. These might be chosen based on phonological behaviour, allowing us to account for allomorphy. Another difficulty to note is the generalisability of parts of the paradigm. By using a neural approach, we wish to leverage the generalisability of the system but to cover even a subsection of language like verbal morphology fully, sometimes a direct exposure to the exceptions is needed.

Ethics Statement

Data on Nen were gathered by Evans under the projects Language and Social Cognition (ANU Aries protocol 2008/253), Languages of Southern New Guinea (ANU Aries protocol 2011/313) and The Wellsprings of Linguistic Diversity (ANU Aries Protocol 2014/224). Nen data are lodged on open access in the PARADISEC archive.

References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and Wholes. Implicative Patterns in Inflectional Paradigms. In *Analogy in Grammar: Form and Acquisition*, page 54–82. Oxford University Press.

Michel Jose Anzanello and Flavio Sanson Fogliatto. 2011. Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics*, 41(5):573–583.

R. Harald Baayen. 2001. *Word Frequencies*, pages 1–38. Springer Netherlands, Dordrecht, The Netherlands.

Louise Baird, Nicholas Evans, and Simon J. Greenhill. 2022. Blowing in the wind: Using 'north wind and the sun' texts to sample phoneme inventories. *Journal of the International Phonetic Association*, 52(3):453–494.

Steven Bird. 2015. Email. *Resource Network for Linguistic Diversity Discussion List*.

James P. Blevins and Juliette Blevins. 2009. *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

James P. Blevins, Petar Milin, and Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In *Perspectives on Morphological Organization*, pages 139 – 158. Brill, Leiden, The Netherlands.

Erwin Chan. 2008. *Structures and distributions in morphology learning*. Ph.D. thesis, University of Pennsylvania, PA,USA.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Nicholas Evans. 2014. Positional verbs in Nen. *Oceanic Linguistics*, 53(2):225–255.

Nicholas Evans. 2015. Valency in Nen. In Andrej Malchukov and Bernard Comrie, editors, *Volume 2 Case Studies from Austronesia, the Pacific, the Americas, and Theoretical Outlook*, pages 1069–1116. De Gruyter Mouton, Berlin, München, Boston.

Nicholas Evans. 2016. Inflection in Nen. In Matthew Baerman, editor, *The Oxford Handbook of Inflection*, pages pages 543–575. Oxford University Press, USA.

Nicholas Evans. 2017. *Quantification in Nen*, pages 571–607. Springer International Publishing, Cham.

Nicholas Evans. 2019. Waiting for the Word: Distributed Deponency and the Semantic Interpretation of Number in the Nen Verb. *Morphological Perspectives. Papers In Honour of Greville G. Corbett*, pages 100–123.

Nicholas Evans. 2020. One thousand and one coconuts: Growing memories in Southern New Guinea. *The Contemporary Pacific*, 32(1):72–96.

Grant Fairbanks. 1960. *Voice and Articulation Drillbook*, Second edition. Harper & Row, New York, NY, USA.

Raphael Finkel and Gregory Stump. 2007. Principal Parts and Morphological Typology. *Morphology*, 17(1):39–75.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the Transformer-based models for NLP tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183.

Nikolaus P Himmelmann. 1998. Documentary and Descriptive Linguistics. *Linguistics*, 36(1):161–196.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of Transformers. *AI Open*, 3:111–132.

Ling Liu and Mans Hulden. 2020. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 153–161, Online. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a Transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Saliha Muradoglu, Nicholas Evans, and Hanna Suominen. 2020. To compress or not to compress? A finite-state approach to Nen verbal morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 207–213, Online. Association for Computational Linguistics.

Saliha Muradoğlu. 2017. *When is enough enough ? A corpus-based study of verb inflection in a morphologically rich language (Nen)*. Masters thesis, The Australian National University.

Saliha Muradoğlu, Nicholas Evans, and Ekaterina Vylomova. 2020. Modelling verbal morphology in Nen. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 43–53, Virtual Workshop. Australasian Language Technology Association.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Andrei Shcherbakov, Saliha Muradoglu, and Ekaterina Vylomova. 2020. Exploring looping effects in RNN-based architectures. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 115–120, Virtual Workshop. Australasian Language Technology Association.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Tom Viering and Marco Loog. 2022. The Shape of Learning Curves: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

A Appendix: Inflection categories

IPFV.FIMP:	Future Imperfective
IPFV.IMP:	Imperfective Imperative
IPFV.MIMP:	Mediated imperative
IPFV.NPHD:	Imperfective Nonprehodiernal
IPFV.YPST:	Imperfective Yesterday Past
IPFV.RMPST:	Imperfective Remote Past
NEUT.PRIM:	Neutral Primordial
NEUT.PRET:	Neutral Preterite
NEUT.PIRR:	Neutral Irrealis
PFV.IMP:	Perfective Imperative
PFV.FUT:	Perfective Future
PFV.PST:	Perfective Past