# Modelling Verbal Morphology in Nen

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE
Australian National University
THE UNIVERSITY OF MELBOURNE

*Saliha Muradoğlu, Nicholas Evans and Ekaterina Vylomova*

## ABSTRACT

Nen verbal morphology is remarkably complex; a transitive verb can take up to 1,740 unique forms. The combined effect of having a large combinatoric space and a low-resource setting amplifies the need for NLP tools. Nen morphology utilises distributed exponence - a non-trivial means of mapping form to meaning. In this paper, we attempt to model Nen verbal morphology using state-of-the-art machine learning models for morphological reinflection. We explore and categorise the types of errors these systems generate. Our results show sensitivity to training data composition; different distributions of verb type yield different accuracies (patterning with E-complexity). We also demonstrate the types of patterns that can be inferred from the training data through the case study of syncretism.

## THE NEN LANGUAGE

Nen is a Papuan language of the Morehead-Maro (or Yam) family, located in the southern part of New Guinea (Evans, 2017). It is spoken in the village of Bimadbn in the Western Province of Papua New Guinea, by approximately 400 people, for which it is a primary language (Evans, 2015, 2020)

- Verbs are the most complex word class in Nen
- Three types of verbs:

| | |
|---|---|
| **Ambifixing:** | Maximal case (up to 1,740 forms for one stem)<br>Employs both prefixes and suffixes |
| **Middle:** | Also ambifixing, but the prefixal slot is restricted |
| **Prefixing:** | Only uses prefixes<br>Separate closed paradigms |

## *Distributed exponence*

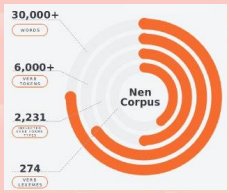
- Marking of grammatical meaning is distributed across smaller pieces of the word, each contribute a subcomponent of that meaning. (Carroll, 2016)
- For Nen:
  - Need to integrate information from prefix and suffix paradigms for TAM, and actor and undergoer

n-ng -owan –t -e
M:α-VEN-set.off-ND:IPF.NP-IPF.NP.2|3SGA
'You/(s)he are/is setting off.'

- In the example above:
  - No one marker marks the singular person
  - -t- marks the subject as non-dual
  - -e marks the subject

## EXPERIMENTAL SETUP

- Muradoglu (2017) provides the verbs from the natural corpus with frequency information.
- This subcorpus is used to generate a set of triplets comprising a lemma, morphosyntactic features, and an inflected form.




- We follow the experimental setup from the SIGMORPHON shared task for reinflection (Cotterell et al., 2016; Vylomova et al., 2020).
- Models: We will utilise two NN models that have shown superior performance in SIGMORPHON–CoNLL 2017 Shared Task:
  - Hard Monotonic Attention (Aharoni and Goldberg, 2017)
  - Neural Transition-based (Makarov and Clematide, 2018)

## EXPERIMENT 2: TRAINING DATA COMPOSITION

***Research Question:*** Does the composition of the training data affect the resultant accuracies, and, if so, how?

- Create training sets according to verb types: prefixing, middle and ambifixing and all combinations
- Each set is made up of 386 instances, limited by the number of prefixing verbs found in the corpus
- Test set is 34 ambifixing, 33 prefixing and 33 middle verbs
- As expected, training sets with one verb type only perform best for that particular verb type
- For the ambi + pre mix, interestingly the results favour prefixing verbs only (likely due to E-complexity)
- For the ambi & middle mix, we would have expected more transfer between both since one is a subset of the other. The difference is likely due to the specific tag used for middle verbs

| | Ambifixing | | Middle | | Prefixing | |
|---|---|---|---|---|---|---|
| | AG | MC | AG | MC | AG | MC |
| Ambifixing only | 11 | 15 | 2 | 0 | 0 | 2 |
| Middle only | 2 | 1 | 12 | 19 | 0 | 1 |
| Prefixing only | 0 | 0 | 0 | 0 | 21 | 24 |
| Ambi + Pre | 1 | 1 | 1 | 0 | 10 | 18 |
| Ambi + Mid | 1 | 4 | 6 | 8 | 0 | 1 |
| Mid + Pre | 0 | 3 | 3 | 10 | 11 | 16 |
| Ambi + Mid + Pre | 0 | 6 | 4 | 8 | 3 | 6 |

Table of absolute number of correct predictions

- The equal split of three verbs set shows significant difference between the A&G model and the M&C

## EXPERIMENT 3: SYNCRETISM TEST

***Research Question:*** Do the models infer properties of the language which are not annotated in the data?

- We test the prediction of an unseen form – 2nd singular past perfective.
- Almost all the TAM categories exhibit syncretism across the 2nd and 3rd singular actor. The past perfective slot is the only case with distinct forms for the 2nd and 3rd person numbers
- The 2nd sg subject past perfective is rare in the natural spoken corpus, with only 2 instances.
- The test set is made up of 100 instances with 98 supplemented from the Nen dictionary
- A&G (2017) architecture incorrectly predicts the 2sg form as 3rd sg (with the suffix {-nd-a} instead of 2sg suffix {-nd–Ø-}) 81 out of the 100 test forms
- M&C (2018) predicts the unseen 2nd sg form as the 3rd singular 90/100 times

## EXPERIMENT 1: TRAINING SIZE & ERROR ANALYSIS

***Research Question:*** How does training size and sampling method affect the models' performance, and what kind of errors are likely across these conditions?

- Training size: HR (10,000 tokens generated by hallucination (Anastasopoulos and Neubig, 2019), ALL is 1,931 (max. available from existing corpus), MR is 1,000 and LR is 100.

| | A&G 2017 | | M&C 2018 | | Non-Neural baseline (NNB) | |
|---|---|---|---|---|---|---|
| | Random | Zipf | Random | Zipf | Random | Zipf |
| HR | 0.610 | | **0.650** | | 0.015 | |
| ALL | 0.390 | | **0.510** | | 0.010 | |
| MR | 0.295 | 0.285 | **0.445** | 0.420 | 0.000 | 0.000 |
| LR | 0.020 | 0.005 | **0.080** | 0.030 | 0.010 | 0.010 |

- We analysed the errors produced in prediction following the taxonomy laid out by Gorman et al. (2019); Di et al. (2019)

| | ALL | | | HR | | | MR | | | LR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A&G | M&C | NNB | A&G | M&C | NNB | A&G | M&C | NNB | A&G | M&C | NNB |
| Allomorphy | 56 | 55 | 190 | 54 | 46 | 144 | 61 | 77 | 188 | 17 | 162 | 190 |
| Free Variation | 30 | 24 | 0 | 14 | 15 | 11 | 13 | 24 | 0 | 0 | 2 | 0 |
| Target | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| Stem | 28 | 11 | 0 | 2 | 1 | 5 | 61 | 7* | 2 | 1749 | 22 | 0 |
| Total | 122 | 98 | 198 | 78 | 70 | 168 | 143 | 116 | 198 | 199 | 194 | 198 |

- Error types:
  - Allomorphy: a misapplication of morphophonological rules, or feature category mappings
  - Free variation: when more than one acceptable inflected form exists
  - Target: mistakes in gold standard
  - Stem: a nonce stem or a re-mapping of a seen but irrelevant stem

## CONCLUSION & CONTRIBUTIONS

- First NN application for the Nen language, and the Papuan language family as far as we know
- We provide a taxonomy of errors produced, with a new category 'Free variation' that arises from the nature of the corpus
- We explore verb type composition effects in training data and the consequent generalisations learnt
- We show that both models learn implicit relationships within the data source (such as syncretism)