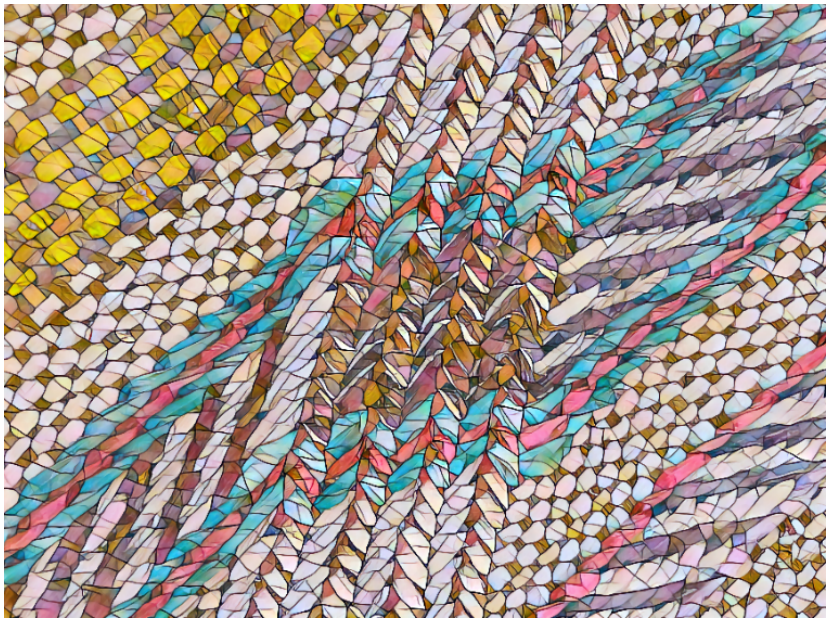
Leveraging computational methods for morphological description: A case study of Nen

Saliha Muradoğlu



Cover: Art adapted from a bag courtesy of the Äkämär dmabende yép Collective.

A thesis submitted for the degree of Doctor of Philosophy

Linguistics

School of Culture, History and Language

Australian National University

January, 2024

Declaration

This dissertation is an account of research undertaken between March 2019 and July 2023 at the School of Culture, History and Languages, The Australian National University, Canberra, Australia.

The work presented in this thesis is that of the candidate alone, except where indicated by due literature reference and acknowledgements in the text. It has not been submitted in whole or in part for any other degree at this or any other university.

This thesis is based on four papers published in refereed scientific venues. My contributions to the research in each chapter and associated papers are detailed at the beginning of each chapter.

Saliha Muradoğlu

22 January 2024

'Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.'

Richard Feynman

Acknowledgements

I want to start this section by sharing a story a wine cellar owner shared with me in Mostar, Bosnia and Herzegovina. This region is mainly known for two wine varieties: *Blatina* (a red) and *Žilavka* (a white). For many years, these two grape varieties and foreign varieties were idolised and thus prioritised for growth and cultivation. An ancient, local grape variety called *Trnjak* was mainly used as a cross-pollinator for the *Blatina*. So neglected was this variety, it was on the brink of extinction 10-15 years ago. Slowly but surely, small local vineyards began recognising *Trnjak* as a winemaker in its own right and focusing its cultivation. It is now a source of local pride, coveted as a rich and complex wine by enthusiasts.

In many ways, this shift of focus from the popular to the little-known, the examination of the old ways to incorporating the new, this story of resilience and perseverance, feels similar to the ethos of this thesis. We are too often blinded by the big and influential, and this serves as a reminder that the understated can have just as much value and meaning. So, I raise a glass in gratitude for all who my path has crossed in this journey. Thank you.

* * *

There are a few major influencers that I want to highlight.

A very special thanks to my family for their endless support, belief, and encouragement, even when I was riddled with doubt. Your combined quest for knowledge, scientific rigour, freedom from fate, autonomy and meaning has led me here. İyi ki varsınız.

A majority of this thesis is based on the frustratingly beautiful Nen language. Unfortunately, I never had the chance to meet any of you in the community, but I am very

grateful to you for sharing your language and culture. I fervently hope I have the opportunity to repay your generosity.

I want to thank my supervisory panel - Nicholas Evans, Danielle Barth, Hanna Suominen, Mans Hulden, and Janet Wiles – for their insights, guidance, and inspiration.

My primary supervisor Nicholas Evans for allowing me to undertake a PhD and for allowing that process to be as a "Free-range" PhD. I took leaps, followed my instincts, had some wins, and made some mistakes, but I can safely say it was an experience I will carry with me. Danielle Barth for helping forge what a thesis by compilation looks like in linguistics at CAP, and for practical advice on all things PhD. Hanna Suominen for her tireless support, encouragement, and sense of rightness, in particular when navigating the complex mechanism that is university administration.

Mans Hulden for being an intellectual inspiration. From the first I read a few papers, I was drawn to this way of thinking and considered it a gold standard to strive towards. Thanks to the Centre of Excellence for the Dynamics of Language (CoEDL) mobility grant, I had the opportunity to visit and meet in person. Despite Covid-19 hindering my intended stay, I have been fortunate to continue our collaboration. I have learnt a lot. It has shaped my thinking and research style. I am eternally grateful, and I hope to pay the favour forward someday.

I extend my deep gratitude to the Australian National University (ANU), the Stephen & Helen Wurm PhD scholarship and CoEDL for being my academic home and support through this journey.

I have to thank the lunchtime ANU sports crew. These guys have put up with all the banter I put out – despite the therapy bills. You guys made the gym a haven. I have witnessed your continued dedication to strength and fitness through the years. You have led by example and are living proof of the importance of generational, cultural transmission. Without you guys, the gym would have been a room with metal and plates.

I'm also very grateful to Daan Van Esch, Ekaterina Vylomova, Huade Huang, James Gray and Mark T. Ellison.

Last but not least, I want to thank my thesis examiners for their valuable feedback which helped shaped this work into its final form.

Abstract

Corpus building is a resource-hungry venture. Collecting, analysing and maintaining are essential stages of corpus building. Modern technology has allowed for collection to be more accessible and, to an extent, has eased the analysis cost. Nevertheless, with the increasing rate at which data can be captured, the resource demands for analysis still need to be lowered. Language documentation poses an acute instantiation of this problem. The extremely limited language descriptions (if any) that we have of most of the world's 7000+ languages, and the needs for community access, only add to the difficulty of sufficiently capturing a language. Further still is the case of endangered languages, where this issue is compounded by time pressure.

This interdisciplinary thesis aims to catalogue the process of enlisting the aid of computational methods for morphological description. The Nen language is used as a motivating case study. Nen is a member of the Morehead-Yam language family of Southern New Guinea. It is an under-resourced language that is actively being described and documented. It provides an interesting case study as it exhibits distributed exponence - a non-trivial means of mapping form to meaning. Its immense morphological complexity and ongoing documentation status make it an excellent playground for providing a real-time account of the process.

In an almost pedagogical order, the thesis presents the development of computational resources for Nen and describes issues, caveats, and opportunities for applying existing computational methods to language documentation. The thesis employs two infrastructures for morphological analysis, finite-state automata and supervised machine learning systems.

The thesis presents the first morphological analyser of Nen and explores the difference between linguist and computational needs. It finds that where a linguist

might consider a principal parts approach optimal for description, computational performance is impartial to a principal part or 'Chunking' approach. The resultant architecture is different, but the accuracies achieved are the same.

The thesis extends its morphological models of Nen verbs with state-of-the-art neural architecture. Its results show no significant difference between random and Zipfian sampling methods, and minor differences may be attributed to the training set composition differences. It introduces empirical evidence highlighting training data composition's effects on model performance. The errors generated by each system are extensively analysed. The most common error types are found to be allomorphy errors, misapplication of morphophonological rules or feature category mappings. Furthermore, the thesis findings indicate that the model learns paradigmatic information as well as string transduction. It predicts syncretism in an unseen form where the rest of the paradigm exhibits syncretism.

The thesis explores the concept of paradigm coverage, a consideration born from the needs of language documentation. The question, '*how much data is enough to capture a language*' has haunted documentary linguistics for decades. The thesis presents another way to measure completeness by leveraging computational models as another line of inquiry; given the existing corpus, how much of the paradigm can a deep learning model capture correctly? It distinguishes between attestation and heuristic coverage; the former describes the corpus presence of specific lemmata and inflectional features, while the latter refers to the discovery process of a morphological paradigm.

Finally, the thesis proposes using active learning to minimise the data costs by prioritising examples where the model showcases difficulty. To explore the viability of this strategy, it extends its subject of study to 30 linguistically diverse languages. It finds that data selection based on model confidence/entropy improves model performance more rapidly than random selection. Experiments included in the thesis imply that these metrics are robust to language typology, with the same behaviour

observed across 30 languages.

Keywords: Nen, computational linguistics, deep learning, morphology, language documentation

Contents

Abbreviations

AL	**A**ctive **L**earning
ANU	**A**ustralian **N**ational **U**niversity
ASR	**A**utomated **S**peech **R**ecognition
CL	**C**omputational **L**inguistics
CRF	**C**onditional **R**andom **F**ield
EV	**E**katerina **V**ylomova
FST	**F**inite **S**tate **T**ransducer
HMM	**H**idden **M**arkov **M**odel
HS	**H**anna **S**uominen
IGT	**I**nterlinear **G**lossed **T**ext
MH	**M**ans **H**ulden
MSD	**M**orpho- **S**yntactic **D**escription
NE	**N**icholas **E**vans
NLP	**N**atural **L**anguage **P**rocessing

PCFC **P**aradigm **C**ell **F**illing **P**roblem

POS **P**art **O**f **S**peech

RNN **R**ecurrent **N**eural **N**etwork

SM **S**aliha **M**uradoğlu

SVM **S**upport **V**ector **M**achine

Research-Output

Peer-reviewed Articles:

[Chapter 2] 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.

Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, Piyush Mishra, Saliha Muradoglu, Lan Sang, Tyler Short, Sagarika Shreevastava, Elizabeth Spaulding, Testumichi Umada, Beilei Xiang, Changbing Yang and Mans Hulden. 2020. Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars. In Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 162–170, Online. Association for Computational Linguistics.

[Chapter 3] Saliha Muradoglu, 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.

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.

[Chapter 5] Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. How to choose data for morphological inflection. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Chapter 4] Saliha Muradoglu, Hanna Suominen, and Nicholas Evans. 2023. A Quest for Paradigm Coverage: The Story of Nen. In Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 74–85, Dubrovnik, Croatia. Association for Computational Linguistics.

Lisa Beinborn, Koustava Goswami, Saliha Muradoglu, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Edoardo M. Ponti, Ryan Cotterell, and Ekaterina Vylomova. 2023. Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. Association for Computational Linguistics, Dubrovnik, Croatia, edition.

Saliha Muradoglu and Mans Hulden. 2023. Do transformer models do phonology like a linguist?. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8529–8537, Toronto, Canada. Association for Computational Linguistics.

Other research contributions:

Data administration for Evans, Nicholas. 2019. Nen dictionary. Dictionaria 8. 1-5005. DOI: 10.5281/zenodo.5526459.

Where applicable code can be found at Github.

Picture: Art adapted from hand-made needlepoint pattern made by Zehra Hasmaden.

Atalarımız sayesinde…

Thanks to our ancestors…

Picture: Art adapted from hand-made crochet pattern made by Saliha Muradoğlu (Sr).

Chapter 1

Introduction

The advancement of modern technology has uncovered innovative ways to make large-scale data collection more accessible and efficient. Driven by the rise of empirical approaches[1], the focus on data has permeated across the scientific community, with data being more important than ever before. This has created a new wave of burden for data analytics. As a result, methods for processing data (including cleansing) and automated analysis have been of increasing interest. While this task remains pertinent across many disciplines, the subject of this thesis focuses on its linguistic manifestation, with particular concern for efforts to document the world's 7000+[2] languages. Language documentation entails developing corpora and grammar for the languages of the world. The documentation effort came to be imperative with half the world's languages predicted to become extinct within this century (Krauss, 1992; Wurm, 2001; Bianco, 2002; Crystal, 2002; Seifart et al., 2018) [3] and of these, around 35–42% are still substantially undocumented (Austin and Sallabank, 2011; Seifart et al., 2018).

[1] For further discussion see Jurafsky and Martin (2009).

[2] For current figure see https://www.ethnologue.com/.

[3] See Evans (2022) for further discussion on calculations and Simons and Lewis (2013) for a break down of language decay by geographic regions.

Languages need not be considered endangered to be of interest. With greater global accessibility, the language sciences have shifted their traditionally Eurocentric domain to a global perspective (Gil, 2001; Evans and Levinson, 2009). Each language contributes an integral part of the puzzle of understanding the full design space of cultural and intellectual diversity. Within each piece lies the opportunity to test, falsify and strengthen the models we develop. For this very reason, the natural language processing (NLP) and computational linguistics (CL) communities have both expanded their application domain to include more languages. The benefits of diversifying subject languages include societal impact (particularly in the case of endangered languages, by creating better language resources), challenging both our computational models and linguistic theories with a wealth of linguistic features, and minimising the risk of overfitting (Ruder, 2020).

One of the primary advantages purported by machine learning approaches to NLP and CL is language independence, intended as a templatic approach to scalability, much like generalisation and abstraction afforded by mathematical thinking. Generalisation is the derivation or induction from something particular to something general by examining the features in common and expanding their domains of validity (Dreyfus, 1991). A notable example of this is ordinary differential equations. This mathematical abstraction is the underlying structure for Newtonian mechanics, reaction rates in chemistry, population modelling and many other natural phenomena (Faraoni, 2020) [4]. It is tempting to generalise in a similar vein across languages with models built on a subset of languages with sufficient resources and theoretically extend to other languages. This would allow for the rapid development of technological resources for smaller language communities.

Yet it is important to note that in the parallel drawn above, the natural phenomena modelled crucially exhibit behaviour defined by the properties of the equation. In other words, the mathematical description is considered an apt generalisation as it

[4] Or indeed partial differential equations (with more than one variable) for sound, heat, diffusion, electrodynamics, thermodynamics, general relativity, and quantum mechanics.

shares properties with the system in question and is, therefore, able to accurately model the behaviour of the system observed at any point.

In the case of language, it does not follow that all of the world's languages must share the same properties and behaviour. While they may share universal features at an abstract level, the surface structures observed in real-world settings (like text and speech) can vary significantly (Ponti et al., 2019).

Most current computational approaches are far from being language agnostic. In fact, often, these methods inadvertently incorporate language-specific biases. More specifically, generalising linguistic phenomena based largely on Indo-European languages yields a lop-sided modelling capacity and performance. As a consequence, discussions in NLP and CL have followed in the footsteps of the language sciences by highlighting the consequences of European-centric practices and focusing on linguistic diversity.

A notable example of this in the NLP and CL domain is language modelling. Language modelling concerns building a probabilistic distribution across a sequence of words and predicting the next word. For example, if asked to complete the following sentence: *'It was the best of times, it was the worst of . . . '*, those who are familiar with the works of Charles Dickens would produce *'times'* with a strong likelihood, but other options such as *'luck'* and *'weather'* are possible though with a lower probability.

Works by Khudanpur (2006) and Bender (2011) detail the case of n-gram language models, which notably underperform for languages with complex morphology (loosely defined as languages with more morphemes per word) and relatively free word order. The n-gram language model is the simplest model used to assign probabilities to a sequence of words. It is predicated on the assumption that the probability of the upcoming word depends only on n-1 previous words (Markov assumption), where n defines the size of the search window. In a bigram model, n is two. In a trigram model, n is three and so on. Similarly, more recent language modelling techniques involving neural works have shown that typological properties

of language, specifically morphological behaviour, has a significant impact on the performance of language models (Gerz et al., 2018).

Moreover, Joshi et al. (2020) shows that low-resourced languages continue to be underrepresented in NLP and CL, and the disparities continue to grow. As a case in point, neural models often overlook the complexities of morphologically rich languages (Tsarfaty et al., 2020):

- Vania and Lopez (2017) details a study on tokenisation based on subword units such as word segments, characters, and character n-grams. Their study is conducted across ten languages: Czech, English, Russian, Finnish, Japanese, Turkish, Hebrew, Indonesian and Malaysian. The results show that the subword unit models are most effective in agglutinative languages. They find that subword tokenisation performs poorly for languages with reduplication (Indonesian and Malaysian).
- Following the previous example, Byte pair encoding (a type of subword tokenisation which essentially encodes the most frequent words as single tokens and less frequent, compound words as multiple tokens of their frequent components, allowing for compression.) does not align well with morphology, as shown by Bostrom and Durrett (2020) for English and Japanese.
- Work by Sabrina J. Mielke et al. (2019) shows that across 69 languages from 13 language families, language models have difficulties with languages that have large vocabularies. [5]
- As above, languages with different grammars, word order and syntax present open problems for neural models. Example studies include Ravfogel et al. (2018) for examining agreement in Basque, Ahmad et al. (2019) looking at the impact of word order on dependency parsing across 30 languages and Hu

[5] Language families considered are: Indo-European, Niger-Congo, Mayan, Austronesian, Sino-Tibetan, Quechuan, Afro-Asiatic, Uralic, Creoles, Constructed languages, Austro-Asiatic, Totonacan, Aymaran. See the paper for full details.

et al. (2020) for noting a sizeable gap in performance for syntactic and sentence retrieval tasks across 40 languages.

- Tsarfaty et al. (2020) highlights that pre-trained embeddings, in other words, vector representations generated in another task and leveraged to a new, similar task, do not always easily encode all relevant information for Hebrew.

The studies listed above emphasise the open questions posed to the NLP world by diverse linguistic phenomena exhibited by the world's languages. They present unique challenges for modelling structure and dealing with data sparsity. Exciting solutions have since risen, such as typologically informed modelling (Hu et al., 2020).

In order to address this 'linguistic diversity gap' in NLP, the efforts must be bidirectional. In other words, while NLP practitioners focus on model architectures and tasks[6], linguists are essential to this effort by highlighting the importance of language-specific phenomena and contextualising the cultural context from which the data is created. Le Ferrand et al. (2022) and Bird (2022) highlight the pitfalls of a purely technical exercise and explore potential solutions that are *'ecologically aware'*.

As the quest for the diversity of linguistic representation grows, so does the need to build corpora. To build a corpus, someone or something has to go through the collected recordings (audio or video) and mark, annotate and describe the contents in a desirable format. Corpus annotation (also referred to as tagging or glossing) can be broadly conceptualised as the process of adding linguistic and other information to a corpus (Hovy and Lavid, 2010). The annotation schemes are often guided by conventions such as the Leipzig Glossing Rules (Comrie et al., 2008) and annotation use, tools and physical format considerations (Ide and Pustejovsky, 2017). For languages undergoing description, the corpus-building endeavour typically falls on the linguist documenting the language.

[6]See section 1.2.3 for examples

Austin (2005) details five stages of language documentation. The first step involves recording media of various formats, including audio, video, image, and text. The next step is capture. This stage involves moving analogue materials to the digital domain. The third is analysis. This stage deals with transcription, translation, annotation, and notation of metadata. The fourth stage is archiving. In this step, archival objects are created, including assigning access and usage rights. The last stage is mobilisation, whereby the materials created are published and distributed.

This thesis fits within the third stage: analysis. Given the high resource cost of linguistic transcription and glossing, a common consequence is a silo of resources, with most recordings shelved without accompanying transcription, annotated materials or translations (Lehmann, 2001; Himmelmann et al., 2006; Evans and Hans-Jurgen Sasse, 2007). This process has been described as a 'bottleneck' in the literature through its various stages. Ćavar et al. (2016) estimate transcription to take up to 50–100 times the real-time audio/video. Another study by Durantin et al. (2017) finds a ratio of 40:1 minutes of transcription time to real-time (recording time) and the average time spent by a linguist transcribing per year as 152 hours. As Himmelmann (2018) puts it, *'It is only a minor exaggeration to say that language documentation is all about transcription'*.

To address the transcription stage of the analysis bottleneck (i.e., transforming an acoustic signal such as speech-to-orthographic text or speech-to-text for short), projects like Elpis (Orthographic) (Foley et al., 2018) and Persephone (Phonemic) (Adams et al., 2018) have been developed. The main concern for Elpis is in making the program usable by documenters with minimal computational experience. The pipeline entails a KALDI-based automatic speech recognition system with a user-friendly web-based graphical interface. The goal of Elpis is not complete annotation, which is infeasible with low amounts of data, but instead, a stepping stone that aids fieldworkers to alleviate the transcription bottleneck by speeding up the process through first-pass suggestions and recognition of recurrent sound units. Persephone also focuses on the language documentation context, with a particular focus on tonal

languages.

Beyond transcription, Evans and Hans-Jurgen Sasse (2007) note the importance of meaning, particularly the role of translation, for language archives. They describe various subpar outcomes ranging from incorrect translations to the worst-case scenario where high-quality recordings of a language are preserved without any translation. To even contemplate whole sentence or utterance translation – the subsequent step after transcription according to the process laid out by Austin (2005) — without any morphological analysis would be premature.

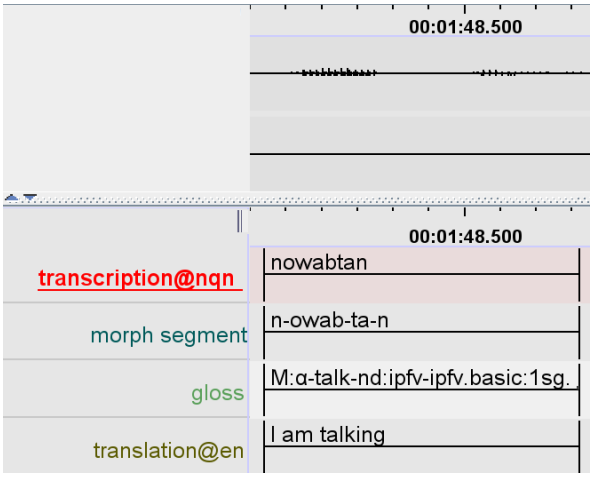


Figure 1.1: Example of an Interlinear glossed text (IGT) in ELAN for Nen. From top to bottom the tiers correspond to a direct transcription of the audio file, a segmented break-down, morphological gloss and an English translation.

Computational models built for translation (machine translation) that utilise neural network architecture are typically trained on parallel corpora[7]. These datasets are comprised of sentence pairs of the two languages studied (incidentally, these correlate to the transcription and translation tier shown in 1.1). An example pair might be the English 'He wrote a letter to a friend' and the Japanese *'tomodachi ni tegami-o kaita'* [8]. To help reduce sparsity encountered in corpora for low-resource languages[9] studies have found benefits to leveraging linguistic annotation. For example, Sennrich and

[7]To address the short supply of parallel corpora, leveraging monolingual data for translation systems is an active area of research (Haddow et al., 2022).

[8]Example adapted from Jurafsky and Martin (2009).

[9]It is difficult to define what makes a language low-resource. 'Resourced-ness' is a continuum, as such any criterion must be arbitrary. It also largely varies with the task in consideration.

Haddow (2016) describes model accuracy improvement when incorporating morphological features, part-of-speech tags, and syntactic dependency labels. Haddow et al. (2022) provides a more general overview of the types of linguistic information utilised to improve translation systems.

For a fully automated pipeline for Austin's stage three, it is possible to conceive of a pipeline using a form of automatic speech recognition for transcription (Elpis, for example), followed by morphological analysis wherein a lexical root and feature values (e.g. speak:PRES.IMPF:1sgSubj) which is then used to generate translations (e.g. 'I speak', 'I am speaking').

After the transcription stage, the order of this pipeline is open to debate. It is plausible to build machine translation models to produce the translation from the transcription, followed by morphological analysis of the machine-generated translation. However, in the case of a language actively undergoing documentation, coupled with the benefits of linguistic annotation to machine translation, the sequence of transcription, glossing and translation is preferable. Therefore, the work presented here focuses on the annotation stage and, more specifically, morphological glossing.

Similar to transcription, manual production of Inter-linearlised glossed text (IGT) takes time and requires linguistic expertise. An idealised[10] depiction of an IGT is shown in 1.1. Minimally, an IGT has the transcription, gloss (i.e. some level of linguistic analysis) and translation tier. Although no estimation for time requirements is currently known (presumably due to the variability across languages), the high cost is well noted. IGTs are essential for linguists, creating language learning materials, other cultural materials and archives, and valuable for many NLP tasks.

Morphological parsing and interlinear glossing are essential for linguists to analyse and document language. Often linguists enlist the use of software tools such as ELAN, Praat and FLeX (Sloetjes and Wittenburg, 2008; Boersma and Weenink, 2018;

[10]The level of detail varies widely depending on the language and the preferences of the linguist/data administrator.

Black and Simons, 2006). These tools range from allowing tier-like structures for annotation (transcription, translation or otherwise), formant and tone analysis, and building a lexicon or basic corpus statistics. The focus of these tools is to be as user-friendly to linguists as possible, which often comes at the price of being heavyweight, rigid, and unfriendly NLP researchers. Although they prove helpful in the initial documentation process, as the corpus grows, so do the research questions and the types of analyses needed. Given the diverse application of this software, it is no surprise that they could be better at addressing a particular stage of the documentation pipeline. Additionally, they rely heavily on software maintenance which can mean a short lifecycle when distributors are susceptible to funding and time restraints. This can create issues with version control. Migrating data from one platform to another can tie up a lot of researcher time. One such example is FLeX and its predecessors Shoebox/Toolbox.

Both ELAN and FLeX can be used for automatic morpheme segmentation and glossing. To utilise such parsing facility, both software require hand implementation of morphological rules. The main drawback to relying on either program for morphological analysis is the lack of predictive power. While FLeX provides a form of scalability by copying the pre-defined segment and glosses, it is vastly limited as it requires the word forms to be identical (Moeller, 2021).

Prior studies have explored automating the task of morphological analysis with computational methods. Snoek et al. (2014) use a rule-based approach (Finite-State Transducers) to obtain glosses for Plains Cree, an Algonquian language. Samardžić et al. (2015) explore automatic IGT glossing for Chintang, a polysynthetic Kiranti language of Nepal. This approach divides the task into two stages: grammatical and lexical glossing. Grammatical glossing is treated as a part-of-speech tagging task, handled by a supervised learning approach, and the lexical glossing is generated using a dictionary. Supervised machine learning requires a set of input observations and the associated outcome (a 'supervision signal'), from which the algorithm learns

how to map from input to output.

Works by Moeller and Hulden (2018), Anastasopoulos et al. (2018) and Zhao et al. (2020) have formulated the problem as a supervised tagging task and used sequence models [11] Moeller and Hulden (2018) consider Lezgian (Nakh-Daghestanian family) for automatic IGT gloss generation. They present a study that compares several methods: conditional random field (CRF) (a sequence classifier that is context aware by calculating conditional probabilities)[12], CRF combined with support vector machines (SVM)(a type of supervised learning which can be used for both classification and regression) and an LSTM-based (Long Short-Term Memory)[13] sequence-to-sequence neural network. The best results are obtained with a CRF model that leverages POS tags. Barriga Martínez et al. (2021) details the development of models for automatic IGT glossing for Otomi, an Oto-Manguean language. The model morphologically segments each word and provides glosses for each segment. Similar to Moeller and Hulden (2018), Barriga Martínez et al. (2021) compares several sequence models, namely CRF, Hidden Markov Models (HMMs)[14] and two recurrent neural networks (RNN)[15] architectures. Again, they found that the CRFs had the best performance.

In Anastasopoulos et al. (2018), they make use of neural network[16] based models with dual sources (Griko and Italian). In a similar vein, McMillan-Major (2020) exploit parallel information (morphologically segmented source language phrase and its English translation) in gloss generation for Abui (Alor–Pantar), Chintang (Kiranti), and Matsigenka (Arawakan). Zhao et al. (2020) furthers this approach by

[11] This describes a set of models whose input have sequential dependence, as is observed in language. By extension, sequence-to-sequence (seq2seq) models further specify the output as sequential as well.

[12] See Chapter 8 in Jurafsky and Martin (2009) for more details.

[13] A type of recurrent neural network (RNN) capable of learning long-distance dependency in sequence prediction problems. See Yu et al. (2019) for overview or Chapter 9 in Jurafsky and Martin (2009) for an introduction.

[14] A Hidden Markov Model is a probabilistic model used to describe a sequence of observed data from a sequence of hidden states. It is often used when the underlying system or process that generates the observations is unknown.

[15] Any network that contains a recurrent unit within its network nodes. In order words, information travels in loops from layer to layer so that the model's state is influenced by its previous state.

[16] See section 1.2.2 for more details.

enlisting cross-lingual transfer and an output length control mechanism for Arapaho (Algonquian), Lezgi (Nakh-Daghestanian) and Tsez(Nakh-Daghestanian).

Another approach to the task of IGT generation is enlisting an Active Learning (AL) framework. AL entails training a model with annotated data (i. e. supervised). At test time, the model is asked to generate labels, which are corrected/verified by human annotators. The corrections are incorporated back into the model for another round of training. In their studies, Palmer (2009), Baldridge and Palmer (2009) and Palmer et al. (2010) implement such a system for the Mayan language Uspanteko. The authors train a maximum entropy classifier to predict a gloss given a morpheme and a context window of two morphemes before and after the morpheme in question. Baldridge and Palmer (2009) note the importance of annotator expertise. They simulate two levels of annotator expertise; the expert is a native speaker of K'ichee', a closely related Mayan language, and has worked extensively on Uspanteko. The non-expert had no prior experience with Uspanteko and only limited exposure to Mayan languages. The annotation process was more efficient and accurate, with forms selected based on model uncertainty for the expert annotator. In contrast, the non-expert annotator glossed more accurately when presented with random IGT rather than the most uncertain.

The main research questions addressed in this transdisciplinary PhD thesis are as follows: (1) how can existing computational methods be leveraged to aid language documentation and description, (2) are there differences between linguistically and computationally motivated modelling choices, (3) how to quantify the amount of data needed to capture all possible features for a subdomain of language, specifically, verbal morphology using the defined inflectional dimensions set out by existing description, and (4) can computational models help prioritise data collection.

This thesis uses the verbal morphology of the Nen language of Southern New Guinea as a case study to answer some of these questions. An outline of Nen's verbal morphology is given in section 1.1. The focus on Nen as a case study is two-fold.

First, in examining the effectiveness of computational methods, there is nothing more realistic than a natural low-resource language subject to ongoing documentation. Second, the morphological complexity of this language sheds light on the modelling of linguistic phenomena and the further deconvolution of mapping complexity.

This thesis presents tools for modelling Nen verbal morphology in a pedagogical manner. First, a FST is compiled as detailed in Chapter 2. In Chapter 3, the application of two neural network architectures is described. The concerns and considerations are evaluated through a documentary lens for each methodology. For example, the kind of information or data needed, the level of detail necessary, and the quantity required to make these tools feasible and valuable. Chapter 4 explores using model performance to evaluate the comprehensiveness of a corpus, from which a prediction of data quantity for complete coverage can be obtained; one corollary is a more refined analysis of the statistical distribution for each slot of the paradigm and the corresponding model performance, with these results showing that not all slots are equal in data demands. Chapter 5 details the use of active learning to prioritise data collection and labelling. Lastly, Chapter 6 summarises and concludes the thesis.

The contributions of this thesis can be dichotomised into resource building and theoretical. From an application perspective, several computational resources were developed to aid in glossing the Nen corpus, and several form uncertainties in description or data were discovered. From a theoretic perspective, a new metric to measure corpus completeness is proposed, and a model-conscious method for data collection can be integrated into the workflow of the documentation process.

1.1 A sketch of Nen verbal Morphology

Nen is a language of the Morehead-Yam Family of Southern New Guinea. It is spoken in the village of Bimadbn, by approximately 300 people. Most inhabitants are multilingual, typically speaking several of the neighbouring languages. This chapter is synthesized based on description and literature by Nicholas Evans (Evans, 2012; Evans, 2014; Evans, 2015; Evans, 2016; Evans, 2017; Evans, 2019b; Evans, 2020; Evans, n.d.).The orthography for Nen used in this thesis follows that set out by Evans and Miller (2016) (see Table A.2 and A.3). Nen is actively being documented. The information presented here is the most current but is subject to being updated; where the description is partial or limited, it is explicitly noted.

The Nen corpus is made of approximately 8 hours of spoken text or over 30,000 words that were recorded in the field with native speakers. A selection of this corpus is available through the PARADISEC archives. This is filtered to over 6,000 verb instances representing 2,282 forms. The full breakdown of the data used for a majority of this thesis is given in figure 1.2.

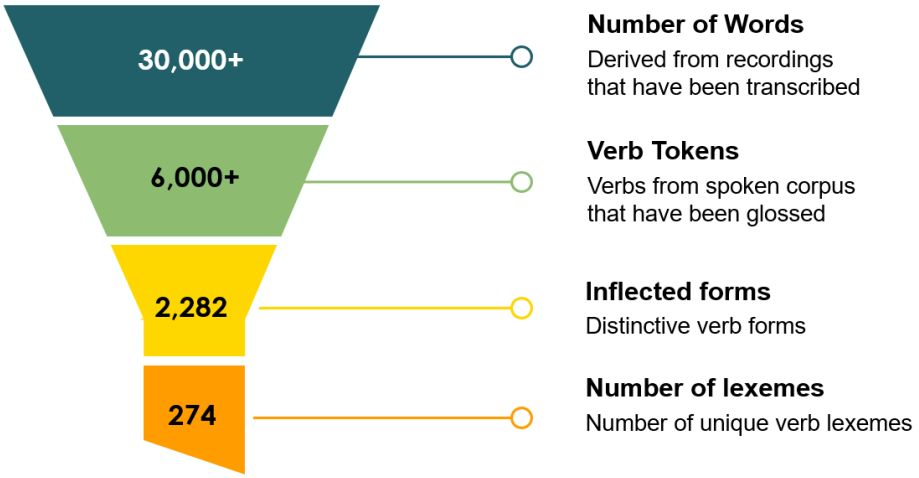


Figure 1.2: The statistics of the current available Nen corpus.

The Nen language showcases complex verbal morphology. In fact, verbs are morphologically, the most complicated word-class in Nen (Evans, 2016). Despite the complexity observed, they are often regular. Figure 1.3 outlines the structure of the

verb for Nen. Verbs inflect for up to two arguments, mark person, number, and a plethora of TAM categories (a more detailed account to follow), as well as direction. Broadly, the Nen verb is made up of prefixes, stem and, suffixes. More specifically, these affixes can be broken down as shown in the verb template in 1.3.

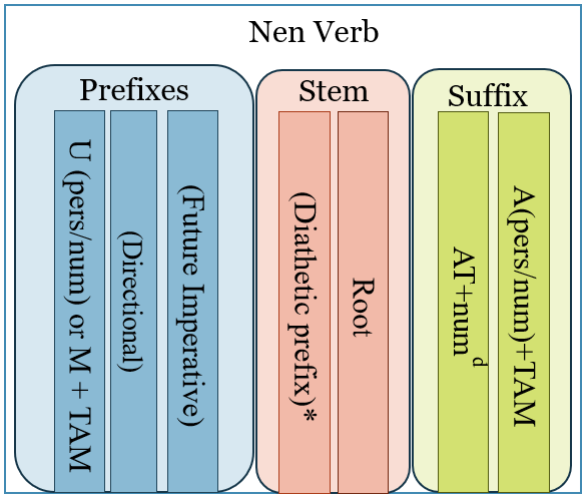


Figure 1.3: A template for the structure of Nen verbs. U(pers/num) refers to the undergoer person/number. M notes the middle verb prefix. TAM refers to the tense, aspect and mood grammatical values available in Nen. AT denotes aspect/tense, num^d refers to the marking of the actor as dual or non-dual, and finally A(pers/num) is shorthand for actor person/number.

In the Nen literature, the obligatory argument indexed by the prefix is referred to as the undergoer and the argument marked by the suffix as the agent. It should be noted that, in using this non-standard use of terminology, the prefix can take a wide range of thematic roles including objects of transitive verbs and indirect objects of ditransitive verbs [17] (Evans, 2016).

Within the prefixal site, the undergoer person/number and tense, aspect and mood (TAM) is marked [18], followed by a directional prefix (marking the action as either towards or away, discussed further in Section 1.1.3). This can be viewed as an

[17] It is important to note the distinction between valency and transitivity. Valency refers to the number of arguments the verb can take in addition to direct objects. While transitivity describes whether a verb can take objects and if so, how many. Prefixing verbs and middle verbs are monovalent. Transitive verbs can be divalent or trivalent (e.g., *räms* 'to give') depending on the lexeme.

[18] Note that this also includes a dummy variable which marks the verb as a middle verb, effectively reducing its valency.

optional prefix or obligatory with a zero-morpheme utilised for an absent directional semantic. When the future imperative semantic is expressed, the appropriate future imperative prefix (marking the actor as either singular *{-ang-}* or non-singular *{-and-}*) must be present. In a similar vein, this can be viewed as optional or indeed encoded with a zero-morpheme when this TAM category is not expressed.

The stem of the Nen verb is comprised of a diathetic prefix and the root. The diathetic prefix encodes various valency alternations (Evans, 2015; Evans, n.d.). These alternations can broadly be split into valency increasing operations (such as causative[19]and benefactive[20], and valency reducing operations (reflexive/reciprocal[21], autobenefactive[22], decausative[23]). The valency-increasing prefix is a w-initial, while the valency decreasing prefixes are vowel initial.

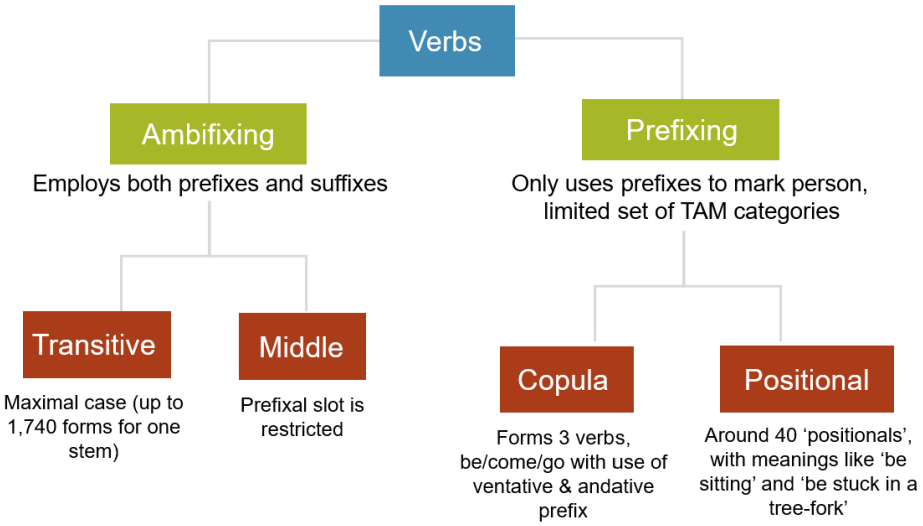


Figure 1.4: A categorisation of verb types in Nen.

[19]These constructions denote the meaning 'cause (motion/trajectory) through sustained contact'. For example, the pair: *armbs* 'ascend', *warmbs* 'take/bring up, cause to ascend'.

[20]These structures carry the 'for the benefit of' semantics. For example, *bens* 'feed', *wabens* 'feed, fatten up (e.g. a pig) for (some future recipient)'.

[21]Structures where each of the participants occupies both the role of agent and patient with respect to the other. An example from Nen is the pair: *wakaes* 'see' and *awakaes* 'see each other/oneself '.

[22]This describes verbs meaning 'for one's own benefit, on one's own account'. For example, *weres* 'listen to', *oweres* 'listen carefully, concentrate on, tune in, attend to'.

[23]These constructions note activity that occurs without overt agent involvement. For example, *nps* 'cut', *eñps* 'get cut'.

The suffix can be further split into the thematic and desinence (See tables 1.5 and 1.6 for paradigm details. The thematic suffix partially marks the tense, aspect, and number (as either dual or non-dual). The desinence notes the actor person/number and TAM.

The mechanism which powers the complexity observed in Nen verbs is described as distributed exponence by Carroll (2016). Distributed exponence is a kind of multiple exponence[24]. Distributed exponence involves the use of more than one morphological segment to convey meaning. It requires all relevant morphs to collectively produce a precise interpretation of the feature value in question Carroll (2016) and Harris (2017). The Nen verbal system distributes information across multiple morphological segments, where the prefixes and suffixes are not independent values. As Evans (2016) puts it affixes *'[need] to be unified into what are in effect circumfixal paradigms before inflectional values are known'*. Consider the following example:

n-n-and-armb-ta-ng

M:α-VEN-FUT.IMP-Nsg-ascend-Ndu:IPFV-NSG.IPFV.IMP

'You | they (>2) climb up later! (in the future, said to a group of people)'

In the example above, to resolve the person/number of the actor, one must narrow options by collecting information from the future imperative prefix, the thematic and desinence suffix. The future imperative marker (discussed further in section 1.1.3) marks the actor as non-singular. Further along the verb, the thematic marks the actor as non-dual (at this point, given the single, dual or plural number distinction in Nen, it is possible to deduce the actor as plural). Finally, the desinence marks the agent as either singular or non-singular (non-singular in this instance). Combining information across all three sites of information, it is possible to mark the actor as plural. The prefix *{n-}* serves as a place-holder variable, marking the verb as a middle

[24]First introduced as extended exponence by Mathews (1974).

	Series		
Person/Number	α	β	γ
1sg	w-	q-	g̃-
2sg	n-	kn-	gn-
3sg	y-/∅	t-	d-
1nsg	yn-	tn-	dn-
2|3nsg	yä-/ya-/e-/i-	tä-/ta-/te-/ti-	dä-/da-/de-/di-
Middle	n-	k-	g-

Table 1.1: Undergoer prefixes, adapted from Evans (2016). The available feature values are contrasted across person/number (noted along the rows) and the dummy variables α, β and γ. These variables are place holders and do not correspond to specific semantic values until they unified with other TAM (Tense, Aspect, and Mood) markings on the verb. The slash notes possible allomorphs.

	Imperfective (IPFV)			Neutral (NEUT)			Perfective (PFV)		
Series	Imperative	Basic	Remote	Primordial	Preterite	Irrealis	Imperative	Future	Past
α	Future Imperative (FIMP)	Non-PreHodiernal, today past (NPHD)	-	Primordial (PRIM)	Preterite (PRET)	Customary /Habitual Past (PIRR)	-	-	Accomplished Past Action (PST)
β	Immediate Imperative (IMP)	Yesterday Past (YPST)	-	-	-	-	Immediate Imperative (IMP)	-	Unexpected Past Action (UPST)*
γ	Mediated Imperative (MIMP)	-	Remote Past (RMPST)	Hope (HP)*	-	Unrealised Action (UA)*	-	Perfective Future (FUT)	-

Table 1.2: Combinations of prefix and suffix series and their meanings. $*$ notes that the TAM category is extremely rare in corpus. Adapted from Evans (2016).

verb (see section 1.1.2). It also marks the verb as a member of the α series. Together with the desinence (and in this case the presence of the future imperative prefix), the TAM feature can be obtained.

While the suffixal system can vary across verb type, the prefix system is shared across verb types. Accordingly, the prefixing system is discussed before delving into the specifics of each verb type listed in 1.4. Table 1.1 shows the undergoer prefixes available in Nen. The undergoer prefixes are divided into arbitrarily labelled series α, β and γ, which do not correspond to specific semantic values until they are considered with other TAM markings on the verb. See table 1.2 for full TAM resolution after cross-referencing with series values. Possible allomorphs are also reported where observed.

In Nen, there are a number of morphophonological alternations observed at the boundary between verb roots and their affixes. For example, when some verbs whose stem ends in final *r* interact with the non-dual imperfective suffix {*-ta*} the expected sequence *rta* is reduced to *na*. Consider the verb *esers* 'to descend', when inflected for the first person-singular basic imperfective ('IPFV.BASIC:1sgA') is *nesnan* 'I descend' (Evans, 2016; Evans, n.d.). This behaviour only affects a subset of verbs with stem final *r*, one notably example which does not exhibit this behaviour is the verb *waprs* 'to make'.

Another alternation observed is vowel harmony. One place this is observed is in some forms of the neutral preterite (Evans, 2016; Evans, n.d.). For example, when the transitive verb *wnḡis* 'to stand up' is inflected for '2|3sgA:NEUT.PRET', the observed form is *ynḡiwi*, instead of *ynḡiwe*. Degemination is also observed. For example, when a stem final *t* is combined with the {*-ta*} suffix the predicted double *t* sequence is reduced to a single *t*, *waets* 'say, tell' forms *yaetan* rather than *yaettan* for 'I say'. The morphophonological variations listed here are by no means exhaustive, others have also been observed. Those presented here are a subset which are straightforward and well-attested. Given the on-going documentation effort, this is an active area of research.

1.1.1 Prefixing Verbs

Prefixing verbs are the smaller set of verbs. Their inflection features are a subset of those available to ambifixing verbs. Specifically, the set of TAM categories available for these verbs exclude the perfective, preterite or irrealis categories (Evans, 2016). Prefixing verbs include the copula *m(n)* (non-dual) *ren* (dual) 'be', the come/go meanings derived by employing the directional prefixes with the copula, *tan*(non-dual)/*wen*(dual) 'to walk' and, approximately 40 'positionals' (discussed further

imminently). With the exception of 'come/go'[25], these verbs lack infinitives.

Copula Verbs

	Tense				
Person/Number	Future Imperative	Nonpast	Near past	Primordial	Far past
1sg		wm	qm	wnzrman	g̃nzron
2sg	nam, gnm	nm	knm		gnnzron
3sg		ym ∼ ymn	tm ∼ tmn	ynzrman	dnzron
1pl		ynm	tnm		dnnzron
2∣3pl	yawam	yäm ∼ yämn	täm ∼ tämn		dänzron
1du		ynren	tnren		dnron ∼ danrwon
2∣3du	yawaren	yären	tären	yärman	därwon ∼ darwon
1pl+					
2∣3pl+	yawamn	yngm, yämn	tngm, tämn		dngnzron

Table 1.3: Nen copula paradigm adapted from Evans (n.d.). From the 'come/go' paradigms primordial forms can be predicted but elicitation of these forms have been unsuccessful so far.

The copula, 'be', paradigm is shown in table 1.3. The 'come/go' paradigms are built using the copula with the addition of directional prefixes, which will be discussed in section 1.1.3.

Positional Verbs

Evans (2014) defines positional verbs as verbs that denote posture (e.g. 'sit' and 'stand') or spatial position in relation to some frame of reference (e.g. 'be up high' and 'be in a tree-fork'). So far 40 verbs of this kind have been documented. Verbs of this class have special stative suffixes *-ngr* for non-dual and *-aran* for dual. As is

[25] A suppletive infinitive *yls* and a special suppletive stem *ewelmän* for large plurals.

		Number			
Tense	**Person**	Sg	Pl	Du	Pl+
Non-prehodiernal	1	w-V-ngr	yn-V-ngr	yn-V-aran	w-V-aran
	2	n-V-ngr	yaw-V-ngr	yaw-V-aran	y-V-aran
	3	y-V-ngr	yaw-V-ngr	yaw-V-aran	y-V-aran
Near past	1	q-V-ngr	tn-V-ngr	tn-V-aran	q-V-aran
	2	kn-V-ngr	taw-V-ngr	taw-V-aran	t-V-aran
	3	t-V-ngr	taw-V-ngr	taw-V-aran	t-V-aran
Remote	1	ḡ-V-ngron	dn-V-ngron	dn-V-aron	ḡ-V-aron
	2	gn-V-ngron	daw-V-ngron	daw-V-aron	dn-V-aron
	3	d-V-ngron	d-V-ngron	daw-V-aron	d-V-aron
Future	2	nang-V-ngr	yong-V-ngr	yong-V-aran	yang-V-aran
Imperative	3	yang-V-ngr	yong-V-ngr	yong-V-aran	yang-V-aran

Table 1.4: Positional Verb paradigm, with V representing the verb root, adapted from Evans (2014)

characteristic of prefixing verbs, positional verbs do not have infinitives and cannot form present imperatives[26].

Table 1.4 shows a generalised paradigm for a positional verb in Nen. *V* represents a positional verb stem. For example, the first-person plural undergoer of the remote-past TAM category '(1plU:RMPST)' for *aki* 'to stand' is *dnakingron*. See Evans (2014) for a list of positional verbs.

1.1.2 Ambifixing Verbs

Middle and transitive verbs share the same TAM paradigm. Infinitives of ambifixing verbs can be identified by the form final *-s*. Tables 1.5 and 1.6 show the available

[26] Distinct from the future imperative, which expresses an imperative to be carried out in the future.

features and their corresponding forms. As in figure 1.3 the thematic needs to be combined with the desinence. For example, to encode neurtal preterite first-person actor, the non-dual thematic *-we-* needs to be concatenated with the desinence *-n*.

	Imperfective			**Neutral**			**Perfective**		
	Imperative	Basic	Remote	Primordial	Preterite	Irrealis	Imperative	Future	Past
Non-Dual	ta	ta	taw	tama	we	nganz	Ø	ng	nd
Dual	e	Ø	Ø	anz	anz	anz	a	a	a

Table 1.5: Canonical forms of thematics across all TAM categories. Adapted from Evans (2016).

	Imperfective			**Neutral**			**Perfective**		
	Imperative	Basic	Remote	Primordial	Preterite	Irrealis	Imperative	Future	Past
1sg		n	n	n	n	n		n	n
1nsg		m	m	m	m	m		m	m
2sg	∅	#e	◀ *e*	nga	∅	∅	∅		∅
3sg		#e	◀ *e*	nga	∅	∅		a	a
2|3pl	ng	t	t	nd	nd	t	ng	∅	t
2|3du	ng	t	t	t	t	t	nd	nd	nd
2|3>du	e	ng	ng	ng	ng	ng	e	e	e

Table 1.6: Canonical forms of desinence across all TAM and person/number categories. Note that # means displacing previous vowel, and ◀ means displacing previous Vw sequence. Adapted from Evans (2016).

Middle Verbs

Middle verbs differ from transitive verbs in valency (Evans, 2015). Another feature of middle verbs is that the stem begins with a vowel. In addition to the diathetic prefix marking the stem as a middle verb, the middle verb is recognisable through the use of a placeholder prefix with no semantic meaning other than to mark the inflection series as α, β and γ and the verb in consideration as a middle verb.

Let us consider the inflected middle verb *nowabtan* 'I am talking'. The first step in analysing this verb is deciding whether it is a transitive or middle verb. Based on the wordform only (i.e., no contextual pronoun such as *ynd* 'I'), the prefix {*n-*} could be '2sg:*α*' or 'M:*α*'. The more revealing component in identifying a middle verb is the stem. A basic strip of affixes yields *owab*. To form the the infinitive a word-final *s* is added, *owabs*[27]. The verb *owabs* is described as a middle verb meaning 'to speak' in Evans (2019a). So this verb can be segmented as *n-owab-ta-n*. Table 1.1 indicates that the prefix {*n-*} marking it as *M:α*. Using information from tables 1.5 and 1.6, it can be deduced that the suffix *-tan* encodes 'IPDV:ND-1sgA' or first-person singular actor, imperfective non-prehodiernal (the basic imperfective for the *α* series). All together:

n-owab-ta-n

M:*α*-talk-ND:IPFV-1sgA

'I am talking'

Transitive Verbs

Extension to transitive from middle verbs is a simple step. Where a limited set of undergoer prefixes are at the disposal of middle verbs, transitive verbs have the full paradigm shown in table 1.1 available (barring the middle prefixes, shown on the last row). Transitive verb utilizes both prefixes and suffixes to mark person and number.

For example, consider *takatang* '(You pl.) look at him/her!'. Again the first step is to identify the stem. Stripping affixes leaves *aka*. Immediately, the prefix prefix {*t-*} can be mapped to the feature value of '3sgU.*β*'. A dictionary search returns *aka* yields a long list of possibilities. To narrow the search a *w* may be inserted to explore the possibility of the stem being transitive. This search (*waka*) returns a much shorter list of verbs: *awakaes* 'see /look at each other, oneself', *wakaes* 'to see/look at', *wakambos*

[27] Another verb *wabs* 'count' looks similar but the lack of 'o' makes it a unlikely stem.

'stalk, creep up on' and *wakaws* 'snatch from'. Examining this list, the closest (in edit distance) verb stems is *akaes* from *wakaes* 'to see/look at'.

An additional piece of information is needed to reach the target form. Evans (n.d.) describes diphthong-shortening for verbs ending with a diphthong for the non-dual thematic , the diphthong is still present for the dual. A quick look at the suffix shows the presence of the non-dual thematic {*-ta*}. So the diphthong 'ae' becomes 'a'. Similar to the example for the middle verb above, information from tables 1.5 and 1.6 is used to decode the corresponding feature values for the suffix. The combined suffix *-tang* encodes 'IPFV:ND-IMP.2plA' or the second-person actor imperfective imperative. Combined together:

t- aka -ta -ng
3sgU.β- see -IPFV:nd -IMP.2plA
'(You pl.) look at him/her!'

1.1.3 Extra combinations

Directional

As seen in figure 1.3 a directional prefix is possible following the undergoer prefix. This can be filled with *{-n-}* 'towards' (ventative), *{-ng-}* 'away' (andative) or left empty to convey a directionally neutral semantic. The direction is relative to the speaker. Consider the verb *armbs* 'to climb', when marked for direction the resultant forms are as follows: *n-armb-te* '(s)he is ascending (neutral)', *n-n-armb-te* '(s)he is coming up (towards speaker)', *n-ng-armb-te* ''s)he is going up (away from speaker)'.

Directional prefixes can combine with a wide range of verbs. Most notable of these is the copula verb. For example, the copula verb *ym* '(s)he is' can become *y-n-m* '(s)he is coming', *y-ng-m* '(s)he is going'. On the surface, these forms are the exact same form as the first singular non-past *ynm* 'we are' and the second/third person large plural

yngm 'they, many, are' of the copula. This homophony does give rise to ambiguity but, is generally resolved based on the surrounding context.

Future Imperative

In addition to normal imperatives, Nen has future imperatives. This type of imperative specifies that an action should be carried out at some later point, and often at a different location (Evans, n.d.). Future imperatives require an additional prefix, which follow the directional prefix (as in figure 1.3). When the future imperative is expressed, the future imperative is marked as either *{-and-}* for a non-singular or *{-ang}* for a singular actor. The future imperative is only possible if the prefix is of the *α* class.

1.2 Intra-word Modelling

Unpacking morphology is often an important step in most NLP tasks (Jurafsky and Martin, 2009; Daniel W. Otter et al., 2021). Before almost any processing of a text, the text must be normalised. This involves tokenisation (i.e., separating words), which can easily be identified by white spaces and lemmatisation for most written languages. Lemmatisation is the process of determining whether two words have the same root. This is obvious in a case like *speak, speaks, speaking* (with *spoke* somewhat trickier), but much less straightforward in the case of suppletive sets like English *am, is, were* and *be*. The most sophisticated methods for lemmatisation involve complete morphological parsing of the word. This step is particularly critical for processing morphologically complex languages.

Rich morphology is more common than one would judge from the computational literature. According to the World Atlas of Language Structure (WALS) (Haspelmath et al., 2005), 80% of the world's languages mark verb tense and 65% mark grammatical case through morphology (Kirov et al., 2018a).

Within traditional NLP, computational morphology can be viewed as a pre-processing step for downstream tasks like machine translation, information retrieval (retrieval of all information on user information need, for example, any search engine) and dependency parsing (analysing grammatical structure in a sentence in terms of dependencies between words). Even in the Transformer era, where entire components of the pipeline described above (normalisation, tokenisation, lemmatisation) is replaced with a Transformer and other non-linguistic sub-word information is used, such as characters, n-grams or byte pair encodings, high-quality morphological processing remains most helpful, especially for morphologically rich languages (Belinkov et al., 2017; Vania et al., 2018; Dehouck and Denis, 2018; Klein and Tsarfaty, 2020; Park et al., 2021; Nzeyimana and Niyongabo Rubungo, 2022).

Within most fieldwork endeavours, IGTs are considered a gold standard for building resources as they are often part of the process when constructing reference grammars (for example a corpus analysis of certain morphological feature distributions), dictionaries (in identifying all unique words), and other language materials (such as storybooks). Morphological modelling is imperative for IGT generation. The analysis involves three stages: segmentation of words into minimal meaningful units (i.e., morphemes), noting the feature values of each morpheme (i.e., glossing) and translation. The implementation of these tasks is discussed further in Section 1.2.3.

1.2.1 Finite-State Transducers

Despite the initially unsettled uptake of finite-state methods in linguistics, led by the influential analysis by Chomsky (1956), it later found purpose in describing subdomains of language. Chomsky (1956) reported on the unsuitability of finite-state languages for describing natural language syntax (*'no finite-state Markov process...can serve as an English grammar'*. Later Chomsky (1959) proposed that four different classes of formal grammars existed in a hierarchy (i.e., the Chomsky hierarchy), including finite-state languages or 'type 3 grammars'. Works by C. D. Johnson

(1972) and Kaplan and Kay (1994) (on phonotactics), and Koskenniemi (1983) (on morphology) demonstrated that certain aspects of natural language are completely analysable using finite-state methods. While, in principle, morphological rules can generate words that cannot be expressed with a finite-state grammar, no such system has been encountered in natural languages (Langendoen, 1981). So, it is no understatement that the hallmark success of computational morphology has been the application of finite-state calculus to morphology.

A finite-state automaton is an abstract computational structure comprising an initial state, transitions, and a set of final states. Pathways from the initial state to some final state define a set of strings (Hulden, 2022). If an input string can be obtained from a path between the initial state and some final state of the automaton, then the string is accepted. If no such path exists, the string is not accepted. A finite-state transducer is an extension of the automaton. FSTs have the same components as an automaton, but each transition is labelled with a pair of input/output strings. These additions allow for both analysis (e.g., `sang` $\rightarrow$ `sing+PST`) and synthesis (e.g., `sing+PST` $\rightarrow$ `sang`).

Since Koskenniemi (1983) combined ideas of sequenced phonological rewrite rules with two-level morphology and proposed a first computational implementation of Finnish morphology, finite-state morphologies have been developed for a diverse group of languages.Prior works include FSTs for agglutinating languages such as Turkish, Tuvan, and Northern Haida (Çöltekin, 2014; Tyers et al., 2016; Lachler et al., 2018), and polysynthetic languages like Arapahoe, Chukchi, Central Siberian Yupik and Kunwinjku (Kazeminejad et al., 2017; Andriyanets and Tyers, 2018; Chen and Schwartz, 2018; Lane and Bird, 2019). The definition of polysynthesis can be nebulous. Greenberg (1960) defines it as something that has a synthetic index (number of morphemes) of greater than three. However, this definition does not allow for distinction between highly agglutinative and polysynthetic morphology. Evans and Hans-Jürgen Sasse (2002) refines this definition as *'a prototypical polysynthetic language is one in which it is possible, in a single word, to use processes of morphological composition*

to encode information about both the predicate and all its arguments...allowing this word to serve as a free-standing utterance without reliance on context.'.

Throughout the history of finite-state approaches, multiple libraries have been developed for ease of implementation (Koskenniemi, 1983; Beesley and Karttunen, 2003; Hulden, 2009; Lindén et al., 2011). With sufficient linguistic insight and investment in developing such models, FSTs allow for analysing any well-formed words in a language. Typically, FST models rely on lexical information, but they can also be used to analyse complex inflections of word forms, provided the morphophonological rules built into the model are obeyed. Even out-of-vocabulary words may be mapped to the closest plausible reading using guessers.

Given their rule-based nature, FSTs have been popular as low-resource languages. In the low-resource language setting, linguistic insight and unformatted materials (such as fieldwork journals) can often be exploited to help generate larger datasets. Moeller et al. (2018) describes using Finite-State methods to produce labelled data for training neural networks. Another study by Beemer et al. (2020) explores the differences between FSTs and neural approaches. In this study, the authors detail the rapid development of 25 grammars for the 2020 SIGMORPHON (Vylomova et al., 2020) shared task for morphological inflection. Most notably, they found that FSTs were only able to outdo the neural counterparts when complex linguistic patterns were observed (such as various inflection classes) and with substantial effort.

1.2.2 Deep Neural Networks

Since the early 1980s, many machine learning techniques have been applied to NLP. These include naïve Bayes[28], k-nearest neighbours[29], HMMs, CRFs, decision trees,

[28] A supervised machine learning algorithm, which is used for classification tasks. It applies Bayes theorem.

[29] A supervised machine learning method, which uses proximity to make predictions about the grouping of an individual data point.

random forests[30], and SVMs (Daniel W Otter et al., 2020). By the end of 2010s, neural networks transformed NLP, enhancing or even replacing earlier techniques (LeCun et al., 2015).

Neural networks[31] remove the need for incorporating detailed knowledge of the specific context by optimising the mapping between input/output pairs. As a consequence, a large amount of training data is required (Gorman and Sproat, 2016). In traditional NLP, features were often hand-crafted, incomplete, and time-consuming to create. NNs can learn multilevel features automatically and have been shown to yield superior results (Young et al., 2018).

To perform higher-level tasks such as translation, text summarisation[32] or text generation[33], understanding of the underlying language is required. This understanding can be broken down into at least four main stages: language modelling, morphology, syntax parsing, and semantics.Computational linguistics has been utilising neural models for various problem sets within these areas. For example, learning supervised morphological inflection (Faruqui et al., 2016; Kann and Schütze, 2016; Makarov and Clematide, 2018b; Aharoni and Goldberg, 2017) and semantic embeddings[34](Mikolov et al., 2013; Pennington et al., 2014) [35]. Most current sequence-to-sequence architectures are based on encoder-decoder models (Sutskever et al., 2014), and usually contain an attention component (Bahdanau et al., 2015; Luong et al., 2015). Encoder-decoder architectures involves a two-stage process. First, the encoder network creates a contextualised, abstract representation from an input sequence. Later, this abstract representation is passed to a decoder network to produce an output that matches

[30]Another machine learning algorithm, used for both classification and regression, which combines the output of multiple decision trees to reach a single result.

[31]Often used interchangeably with deep neural networks. Modern neural networks are often deep (i.e., have many layers).

[32]The task of generating a shorter version of a document while preserving its content. See Lloret and Palomar (2012) for a detailed overview.

[33]The task of producing new text, with the goal of imitating human language patterns. See Celikyilmaz et al. (2020) and Iqbal and Qureshi (2022) for survey of task.

[34]A vector representation of a word, where this vector encodes the meaning of the word in such a way that words that are closer in the vector space are expected to be similar in meaning.

[35]See De Mulder et al. (2015) for an overview on Language modelling

the desired format/task. These properties render encoder-decoder models capable of generating contextually-aware and length-variable output sequences from input sequences (such as sentences for machine translation). These architectures face difficulties with long sequences as the information from the beginning of the input sequence is diluted as the subsequent segments are processed. Inspired by cognitive attention, the *attention component* was propose as a solution to this problem. This mechanism allows for significant parts of the input sequence to be prioritised[36].

To address the increasing concern of diverse linguistic representation, highly multilingual corpora such as UniversalDependencies (Nivre et al., 2016) and UniMorph (Sylak-Glassman et al., 2015; Batsuren et al., 2022b) have recently been introduced. SIGMORPHON organised a number of shared tasks on morphological reinflection starting from 10 languages in 2016 (Cotterell et al., 2016) and up to 90 languages in 2020 (Vylomova et al., 2020). In 2020, languages were sampled from various typologically diverse families: Indo-European, Oto-Manguean, Tungusic, Turkic, Niger-Congo, Bantu, and others. To date, only two Australian Aboriginal languages Murrinh-patha (Vylomova et al., 2020) and Kunwinjku (Pimentel et al., 2021) have ever been included in the shared task, and one Papuan language – Eibela (Pimentel et al., 2021). Eibela comes from the large Trans-New Guinea family and is entirely unrelated to, and typologically very different from Nen.

The neural network approaches adopted in this thesis are supervised approaches. In other words, a labelled dataset is used to *supervise* or guide the model. The primary reason for this is that the intention is to integrate computational methods with existing documentary efforts and analytical choices. Accordingly, the existence of a reasonably stable morphological analysis is assumed. The extent of 'reasonable' may be interpreted generously; mainly, data full of contradictory or nonconforming examples is to be avoided. Nevertheless, the description may be evolving.

[36]See https://zhanghanduo.github.io/post/attention/ for a friendly introduction for sequence-to-sequence models and attention

1.2.3 Task

Within computational morphology, computational tasks can be categorised into two major groups: generation and analysis (Liu, 2021) (akin to the bi-directionality of an FST, allowing for both analysis and synthesis). Within the taxonomy presented by Liu (2021), segmentation is categorised as analysis, while glossing or tagging can be either generation or analysis depending on the input/output pairs. For generation, the input is a stem with morphosyntactic tags. For analysis, an inflected form is given, which is decomposed into its stem and the relevant morphosyntactic tags.

The first step in analysing a word's morphology is to establish morpheme boundaries. In 2022, one of the SIGMORPHON shared tasks focused on morphological segmentation (Batsuren et al., 2022a). Previous studies addressing morpheme segmentation approaches utilise unsupervised methods, where the objective is to induct morphemes (Goldsmith, 2001; Smit et al., 2014; Soricut and Och, 2015; Eskander et al., 2019; Eskander et al., 2020). Segmenting a word into morphemes involves defining morphemes. This is most likely achieved in the earlier stages of documenting and describing a language. In the case of Nen, this description exists (see Section 1.1 for details), so the experiments presented in this thesis focus on the subsequent stage of morphological analysis, that is, glossing and generating inflected forms.

The second step of analysis is attributing morphosyntactic values to each segment – i.e., glossing/tagging. This stage can be conceived as an analysis or generation task depending on the input/output pairs. The focus of experiments presented here is on the generation side, or the task for morphological (re-)inflection as defined by the SIGMORPHON shared tasks (Cotterell et al., 2016; Cotterell et al., 2017; Cotterell et al., 2018; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023).

This choice is motivated by several factors. Firstly, the task of generating a full morphological paradigm based on a lemma allows a linguist or native speaker to

evaluate the predicted forms[37]. This task also works as a diagnostic tool; for example, when the model provides a prediction, and the linguist or native speaker assesses the prediction as incorrect, is this due to the lemma being an exception, the description being insufficient or simply a model inadequacy? If either of the first two is correct, this might suggest further investigation is needed. If the latter option, error analysis can help identify problem inflection patterns for the model, the kind of data needed or in some cases detect errors made by human-annotators[38]. These products of the inflection task make it a good fit for the objective of aiding morphological description. Secondly, the wide range of languages considered in the yearly shared task makes for a good body of literature for comparison. Thirdly, the task design allows for string transduction (the mapping of one word form into another), making it an ideal task for sequence-to-sequence models (Nicolai et al., 2015; Rastogi et al., 2016; Nicolai et al., 2018; Ribeiro et al., 2018; Makarov and Clematide, 2018b; Makarov and Clematide, 2018a; Wu et al., 2018). Lastly, it allows paradigmatic information to be modelled.

In the SIGMORPHON inflection shared task setup, the model is exposed to triplets of the lemma, morphosyntactic descriptions (MSD)and inflected form at training time. For example, a triplet for the English verb *eat* might be ⟨`eat, V;PRS;3;SG, eats`⟩. The model is prompted to generate the inflected form with only the lemma and MSD during testing. Morphological generation can be either type or token-based.

The two most extensive standardised, cross-lingual datasets and schemas for morphological annotation are supplied by the Universal Dependencies (UD) (Nivre et al., 2016; Nivre et al., 2020; Marneffe et al., 2021) and Universal Morphology (UniMorph) (Sylak-Glassman, 2016; Kirov et al., 2018b; A. McCarthy et al., 2020; Batsuren et al., 2022b) projects. Before exploring the differences between these projects, examining the distinctions between types and tokens is important. Consider the following quote:

[37] Eliciting or inflecting for full paradigms is anecdotally more time-consuming than assessing a provided form as correct or incorrect.

[38] See 3for target error examples.

> When life gets tough, sometimes all you need is a tree to talk to.
>
> (José Mauro de Vasconcelos — My Sweet Orange Tree)

How many words does the above sentence contain? One way to answer this question is to count each occurrence of a word[39]. This is referred to as a token. So, the quote above has 14 tokens. Another way to answer this question is to tally all unique tokens. Repeating the word *to* twice means there are 13 unique tokens or types. Counting the number of unique dictionary entries each word represents (i.e., headwords) is also possible (Brezina, 2018). A lemma or citation form is the grammatical form that is used to represent a lexeme (e.g., trees to tree or went to go), so the quotation above has 13 lemmas. A lemma with a particular meaning attached to it is referred to as a lexeme. This distinction arises from the need to distinguish polysemous words. In the quotation above, the two uses of *to* are different senses of the word; the first is to mark the verb talk as an infinitive, and the second as a preposition to indicate the target or recipient of an action. Thus, this quote has 14 lexemes.

As mentioned above, the task of morphological generation can be implemented as a type or token-based task. Accordingly, two projects address this distinction. The UniMorph project details an annotation schema that aims to be cross-lingual and is a type-based resource. A type-based morphological generation task concerns generating inflected forms from a lexicon. In other words, the data contains a lemma in isolation, inflectional values and the corresponding wordform. By contrast, the UD project is a token-based annotation scheme which considers context – typically at the sentence level. The dataset is comprised of annotated treebanks[40]. Since both schemata encode similar information, it is possible that both projects may be needed

[39]The definition of a word can be complicated, even more so in languages such as Chinese, Japanese, Arabic and Hebrew where word segmentation is notoriously more difficult as there are no explicit word boundary markers (such as whitespaces) Xue (2003) and Shao et al. (2018). Here, the most basic definition is used – segments demarcated by whitespace.

[40]Treebanks are parsed corpora that are annotated for syntactic structure (in the case of UD, these structures are analysed in terms of dependencies (described from dependency grammar) at the sentence level.

at some stage of language description or corpus building. A. McCarthy et al. (2018) outlines a mapping process from one project to the other.

Current resources available for Nen include 30,000+ words of transcribed text, from which all verbs have been hand-annotated to form a specialised dataset Muradoğlu (2017). Unfortunately, annotation involving syntactic or semantic parsing is yet to be done. Taking into account the available data and the aim of aiding morphological description, the work presented here employs a type-based (i.e., following the schema of UniMorph) database for Nen and the setup of the SIGMORPHON shared tasks on inflection.

While the task of (re-)inflection focuses on inflected form generation, more recent work invites a focus on the production of IGTs directly (Ginn et al., 2023). The input for this task is the transcription and translation of a text, and the desired output is the morphological gloss. A linguist might identify and collate inflected forms encountered in natural speech (or indeed a corpus of IGTs) when describing a language. This process is formulated as a separate task by Moeller et al. (2020). This task pertains to the production of complete inflectional paradigm tables from IGTs.

Given that language naturally exhibits a Zipfian distribution (Zipf, 1932; Zipf, 1935), texts collected in the process of language documentation often need to be supplemented to create complete a grammar. For example, to describe the noun morphology of one language, a linguist typically must collect paradigms of these inflection patterns for a large number of lemmata (Moeller, 2021), in doing so in a targeted way since they do not naturally occur with significant frequency in language. Zipf's law[41] describes a mathematical power relation[42], where the measured value (word frequency in Zipf's notable example) is inversely proportional to its frequency rank. For example, the second most frequent work in English *'be'* will be half as frequent as

[41] or in the author's own words *'Principle of Relative Frequency'*.

[42] Despite the notable dislike of mathematical involvement in linguistics. *'Let me say here for the sake of any mathematician who may plan to formulate the ensuing data more exactly, the ability of the highly intense positive to become the highly intense negative, in my opinion, introduces the devil into the formula in the form of* $\sqrt{i}$*.'* Page 21 in Zipf (1932).

the most frequent item *'the'*, and the third most frequent item *'to'* will be a third as frequent (James P Blevins et al., 2017).

One way to address this sparsity can be formulated as the paradigm cell filling problem (PCFP). PCFP has its origins in language acquisition; morphologists have explored the abstraction process of parts to wholes [43]. PCFP, as outlined by James P. Blevins and J. Blevins (2009), asks how speakers of a language can reliably produce inflectional forms of most lexemes without ever witnessing those forms. The basis of this problem is generalisation, how a full paradigm may be constructed from partial ones. Computational implementations include an encoder-decoder approach (Silfverberg and Hulden, 2018), and the CoNLL-SIGMORPHON 2017 Shared Task (Cotterell et al., 2017) where systems were given a lemma and some of its specific inflected forms and asked to complete the inflectional paradigm by predicting all of the remaining inflected forms. By extension, the theoretical question of identifying principal parts and determining the optimal path to generalisation arises. Like a language learner, a field linguist typically obtains part of the paradigm (either through elicitation or natural means) for each word. These fragments likely allow for reconstructing the entire paradigm (Ackerman et al., 2009; Liu and Hulden, 2020). Another way to address this sparsity[44] in representation is by artificially curating the data on which the model is trained, which will be explored further in Chapter 5.

1.3 A data-conscious perspective

Data is an integral component of empirical research. It encapsulates a concrete way to capture direct and indirect observations to support or challenge a purported hypothesis. This section discusses the considerations of data and its importance in both linguistics and computational modelling.

[43] or the other way around depending on the morphological framework used.

[44] Assuming, of course, that low model confidence mostly correlates with rarity. This is explored further in Chapter 4.

In linguistics, a major discussion is how data should be collected and collated. The field of corpus linguistics, which involves the study of language through a corpus, exerts a substantial amount of focus on corpus design. Building a corpus is not merely a task of collecting anything and everything. In the design stage, several decisions need to be made. These decisions include but are not limited to how much data is needed to minimally achieve the goal of the project and what kind of annotation is needed (e.g., should part of speech be marked or is it irrelevant, should the transcription mark stress) (O'Keeffe and M. McCarthy, 2021).

The broader goal of the research project impacts the decisions described above. For example, Barth and Schnell (2021) describe three types of corpora: general, language documentation and research corpus[45]. The goal of a language documentation project is to create linguistic descriptions, dictionaries and other non-linguistic tasks (such as resources for the language community). The corpus of such a project is typically small, comprised of mostly spoken recordings, linked to primary media and requires processing/annotation. By contrast, the research corpus (type 3) is created with a focused research topic in mind. As Biber (1993) describes it, *'there must be a match between the language being examined and the type of material being collected'*.

1.3.1 Corpus Representativeness

The use of corpus data for linguistic investigation relies on one central assumption, namely that the corpus is representative of the linguistic phenomenon in question (Raineri and Debras, 2019). A corpus is a sample of a language or language variety (i.e. population) (McEnery et al., 2006). Leech (1991) characterises a corpus as representative of a language variety when findings based on its contents can be generalised to a larger hypothetical corpus[46]. Similarly, Biber (1993) defines representativeness as 'the extent to which a sample includes the full range of variability in a population'; it is a subset of a population that aims to reflect the characteristics of the larger group.

[45] See Table 6.1 on page 92 in (Barth and Schnell, 2021) for full characterisation.
[46] See Kruskal and Mosteller (1980) for a historical overview on representative sampling.

When constructing a corpus, Biber (1993) notes that various factors such as genre, register, and medium (i.e., spoken or written) must be considered to provide adequate coverage of a language at a given time and domain. A corpus's overall design encodes decisions about the types, quantity, and length of texts to include. Each of these decisions involves a sampling decision (Biber, 1993; Biber et al., 1998). The length of text and number of texts per variety pertain to the balance of the corpus. That is, the weighting given to different elements in a corpus. For example, if more legal texts are included, the distribution of linguistic phenomena in the corpus might be more representative of legalese rather than linguistic structures found in narrative texts.

Another thread of consideration for balance and representativeness in corpus design is whether it should be bottom-up or top-to-bottom; in other words, whether texts should be curated based on internal or external criteria[47]. Internal criteria concern the distribution of words or grammatical features; by contrast, text-external criteria are based on parameters related to the context of its production, such as its author, the intended aims of the author or perception by the audience (Sharoff, 2017). Examples of external criteria are genre and register. McEnery et al. (2006) highlight the circular nature of using internal criteria as primary metrics for selecting corpus data. One primary function of a corpus is to study linguistic distributions. If, during the design phase, the corpus is curated based on internal criteria, any resulting study does not reflect the natural feature distributions of the language. The corpus has been skewed by design. Sinclair (1995) suggests text selection based on external criteria in the initial phase of building a corpus. After which, corpus analysis can guide the representativeness of the corpus. Biber (1993) suggests that *'the compilation of a representative corpus should proceed in a cyclical fashion'*.

A fundamental consideration for fieldworkers is providing the best possible account for a language with limited time and resources. Where should they spend their resources? Evans (2008) outlines the debate around whether the primary focus for

[47] In Biber's literature, this is referred to as 'linguistic perspectives' and 'situational variability' respectively.

capturing language should be expanding the corpus of primary data or developing a description. The review details two cases to support the supplementation of unstructured text with elicited probing. The first case presented by Hyman (2007) centres on tone in a Kuki-Chin language, Thlantlang Lai. The author argues that elicitation is needed to work through all possible tone combinations. The second case by Rice (2006) highlights a range of principles uncovered through the use of elicitation to understand how Slave syntax works. Both cases draw attention to the fact that even an infinitely large corpus is unlikely to contain all combinations needed to answer questions about specific phenomena. Often, corpora constructed during documentation efforts are a combination of both.

The language documentation case is a particular instantiation of the sampling problem; if the population is still being discovered, how might you even begin to measure the corpus/sample representativity? In addition to this, Atkins et al. (1992) notes that even if it were possible to delimit the population in a rigorous way, given resource limits, finding an inadequately represented feature in the sample would always be possible. Given the mix of primary data and targeted (often elicited) recordings, how representative is this of the language studied?

In most language documentation projects, it is important to note that the field recordings, which often directly contribute to the language corpus, are directed by the heuristics of the linguist doing the fieldwork[48]. In a sense, this is another type of population sampling. The high variability of individual intuition is unavoidable, given the nature of fieldwork. Leech (1991) describes representativeness in corpus linguistics as an *'assumption [which] must be regarded largely as an act of faith'*. This is arguably even more true for language documentation corpora.

Although it is improbable for a small corpus to contain samples from all environments citep{o2010routledge, that should not discredit insights obtained from smaller corpora. Biber (1990) reports on the relative stability in the occurrence of common

[48]Direction of research is often guided by a linguist 'following their nose'.

linguistic features (such as prepositions, past/present tense and personal pronouns) across 1,000-word texts obtained from the London-Lund corpora. While this study focuses on English, it highlights that even a relatively small corpus can sufficiently represent the language in question, provided the corpus represents the full range of variation. In the Nen corpus, the attestation of morphosyntactic features is shaped by the availability of texts and their lengths. A clear example of this is the appearance of two inflected forms of the transitive verb *yis* 'to plant' in the top 20 most frequent verbs across the corpus[49]. This frequency is likely to be boosted by the many coconut interviews[50]. Nevertheless, the most frequent verb is the copula verb, occupying half of the top 20 most frequent verbs. Other verbs include *räms* 'to do/give', *owabs* 'to speak' and *wakaes* 'to see'. All these verbs can be used in more widespread contexts and are not as limited in the range of us as *yis* 'to plant'.

The concept underpinning the Nen *yis* 'to plant' spike was first introduced by (Kilgarriff, 1997) as the whelk problem. Suppose an English corpus contained a substantial sample from a book on whelks; the word whelks will appear frequently given the subject content. In general English (i.e., the population), whelk is not a particularly common word. The problem arises when a frequency list is calculated from this corpus; the term whelk will appear to be high frequency. In other words, the sample is not representative of the population. To counter this, an additional metric -— dispersion is introduced. Dispersion accounts for the distribution of words or phrases throughout the corpus. Various metrics for dispersion exist; for an overview, see Gries (2008). Considering the growing nature of a language documentation corpus, iterative corpus analysis to guide representativeness and balance is ideal. It must also be tempered with the practicalities of fieldwork, for example, working with what stories exist and what community members wish to share. Adding a more

[49]The ninth most frequent wordform is *yiwin* - the *yis* verb inflected for first person-singular neutral preterite ('NEUT.PRET:1sgA') and the eighteenth most frequent *yiwi*, *yis* inflected for second/third person-singular neutral preterite ('NEUT.PRET:2|3sgA').

[50]These typically involve some biographical questions and questions about coconut trees that belong to the interviewee. See Evans (2020) for more details.

comprehensive range of genres and registers remains an ongoing effort for the Nen corpus.

1.3.2 Data Quantity

As mentioned above, the quantity of data is an important consideration when constructing a corpus. Given that one of the goals of language documentation is to describe the linguistic design space, a further consideration is how to quantify the amount of data points needed to obtain a complete outline of the system in consideration. For example, Baird et al. (2022) explore the question of how much text is needed[51] to capture the inventory of phonemes[52]. The authors count the phoneme attested when at least one allophone is encountered.

Another example is from morphological paradigm attestation. Muradoğlu (2017) attempts to answer the question of how much data is enough to have full verbal paradigm attestation for Nen. The study examines the frequency distributions of Nen verbal paradigms and the coverage of such paradigms (for different verb types) within a 30,000-word corpus, of which more than 6,000 were verbs. Like English, the most frequent verb in Nen is the copula *ym* ('3sg:NPst' 'be'), followed by the same root inflected for different TAM categories (*dnzron* '3sg:farPst',*wm* '1sgNPst', *yngm* '2|3pl+:NPst' or '3sg:AND:NPst' meaning '(s)he comes'.). The frequency distributions reported were Zipfian. Zipf's law has been shown to hold across various languages (Piantadosi, 2014) for word frequency and other units of language such as phonemes (Macklin-Cordes and Round, 2020). This poses a problem for any documentation project aiming at comprehensive coverage since many of the key items (e.g., some cells in large paradigms) will only occur very infrequently and hence only occur once a very large corpus is obtained.

[51]The authors formulate this as the 'Himmelman-Bird' problem, which is discussed further in Chapter 4.

[52]The smallest unit of sound which distinguishes meaning.

An example is seen in Nen, where a transitive verb can take up to 1,740 distinct forms. The calculation of this daunting figure involves summing across the forms associated with each series (α, β and γ). For each series, all possible TAM and actor(person/number) combinations are counted[53]. The number obtained is then multiplied by the possible undergoer prefixes[54]. Certain combinations are disregarded as they encode semantics which are expressed through different constructions. For example, a first singular actor acting on a first singular undergoer, this semantic is expressed by a reflexive form. Finally, this number is multiplied by three for each direction prefix available (see section 1.1.3). Muradoğlu (2017) reports a 3.06% coverage for *räms* 'to do/give', the most frequent transitive verb.

Given that middle and transitive verbs in Nen share the same actor/TAM inflectional paradigm, it is conceivable that most of the verbal inflection system can be constructed from either of the prefixing verbs (copula or positionals) and the middle verb paradigm. When the prefixes from a prefixing verb and the suffixes from the middle verb are combined, most of the transitive verb paradigm is recovered. The only additional information needed is the *interactional* forms (e.g., '2|3sg du' for the imperfective series, *-ng*). By taking the minimal set of inflected forms of a lexeme which allow for all other possible inflections to be deduced (i.e., a principal parts approach (Finkel and Stump, 2007)), the coverage increase to approximately 95% and consequently the estimated quantity of data needed is reduced (Muradoğlu, 2017).

The answer to the question of the quantity of text/data needed largely depends on the data sample considered. Baird et al. (2022) considers the 'North Wind and the Sun' texts, while Muradoğlu (2017) considers the Nen corpus of approximately 30,000 words made up from naturalistic texts. How can the data required be minimised? For English, the 'Rainbow Passage' by Fairbanks (1960) captures all phonemes in the first four lines. This idea can be formalised as the Kolmogorov complexity (Kolmogorov,

[53]Note the syncretism of several cells. For example, the suffix *{-te}* encodes the imperfective basic and remote second and third person singular (i.e., 'IPFV.BASIC2|3SGA and IPFV.RMPPST:2|3SGA.

[54]This is where the calculation for paradigm size differs for middle verbs, as they only have three prefixes available. See table 1.1.

1963). Kolmogorov complexity is the shortest computer program (or set of rules) that produces the desired outcome. Although analytically tantalising in its elegance, Kolmogorov complexity is impossible to calculate in practice. Instead, pursuits of optimisation of this kind focus on reducing the required information rather than reaching the absolute minimum. This process describes compression.

A classic linguistics example includes the grammar of Sanskrit written by Pāṇini. This work seeks to extract all grammatical regularities rigorously guided by the twin imperatives of complete coverage and the principle of minimum description length[55]. Pāṇini builds semantic sense with each successive rule so that no word or meaning is repeated in subsequent sentences. Almost every sūtra in the Aṣṭādhyāyī is an elliptical sentence that borrows meaning from the sūtra or sūtras above it. In its entirety, the Aṣṭādhyāyī consists of just 7,007 words. Suppose the grammar where fully decompressed; the rule set expands to 40,000 words. Paṇīni achieves a 1/6 compression (or 1/3 bytes) (Goyal et al., 2007; Chandra, n.d.).

Intuitively, a more complex system is more challenging to describe. For example, a higher-order polynomial requires more data than a linear system to fit a curve (and thus model the system). The concept of complexity is hard to define and has been an ongoing discussion among linguists. Ackerman and Malouf (2013) make a distinction between E(numerative) complexity and I(ntergrative) complexity. E-complexity identifies the number of morphosyntactic feature categories (i.e., paradigm cells) available in a language, while I-complexity measures the predictability among word forms within a language. Other works show a trade-off between the paradigm size and irregularity: a language's inflectional paradigms may be either large or highly irregular, but never both (Cotterell et al., 2019a).

A similar concern on sparsity is noted in computational learning theory[56](Pertsova,

[55]The minimum description length principle states that the best explanation for a set of data is one that allows the most compression of the data. In other words, the explanation that encodes it in the shortest possible way. Ideally, this would be equivalent to the Kolmogorov Complexity of Sanskrit.

[56]An area of research which uses computational modelling to understand further the task of learning a natural language (i.e., language acquisition).

2016). Focusing solely on E-complexity, Chan and Yang (2008) shows the relationship between corpus size, paradigm size and saturation (i.e., how many available forms are attested in the corpus). It is clear from the results that the more E-complexity a language has, the larger the corpus needed to capture it. Lignos et al. (2016) also reports on the data sparsity encountered in child-directed language. With only a small percentage of possible morphological forms attested in acquisition data, the authors conclude that the child learner must be able to form wide-ranging generalisations from partial paradigms.

In machine learning or NLP projects, creating training data is often the most expensive part in terms of cost and time (M. Johnson et al., 2018). As such, predicting the quantity of training data and its relationship with accuracy is of great importance. Plotting learning curves[57] is an established practice for investigating these parameters. The common wisdom for improving accuracy is to acquire more data (Banko and Brill, 2001). However, this is not always feasible, particularly for under-resourced languages. M. Johnson et al. (2018) set out to model how accuracy varies as a function of training size on subsets of data and use that model to predict how much training data would be required to achieve the desired accuracy. Williamson (2020) suggests treating data as a process rather than a thing; it encapsulates decisions that should, in theory, lend themselves to the research question of interest. However, in addressing the research question and the implications for the data collected/used, how the data is substantiated must also be considered.

1.3.3 Data Quality

Another important factor to consider is data quality. Data remains susceptible to errors or noise that may be introduced during the collection, annotation or collation stage (Gupta et al., 2021). For example, noise in morphological inflection data can be in the form of incorrect annotations or mixed orthographies (Gorman et al., 2019;

[57] A learning curve of a machine Learning model shows how the model performance (this metric could be loss, accuracy, error) changes as the size of the training set increases.

Moeller et al., 2020). Liu and Hulden (2022) use the Transformer model to identify annotation errors by artificially generating three categories of errors: typographical (change of a single character in data), linguistic confusion (two inflected forms are systematically swapped) and self-adversarial (where the model itself is used to generate incorrect yet plausible-looking forms). Wiemerslage et al. (2023) investigate the types of noise encountered in inflection data and the impacts for the task of unsupervised morphological paradigm completion.

Data quality is among the most commonly cited culprits for poor model performances. However, the concept of quality is ill-defined [58]. What makes a dataset high-quality? The standard practice for verification involves comparisons to human benchmarks after the training set has been re-evaluated. Quality can be defined as a high signal-to-noise ratio, where noise is misinformation or irrelevant information. With this working definition, high-quality data for the morphological inflection task would have standardised orthography, little to no mislabelled MSDs and consistent labels (e.g., no leftover analyses resulting from ongoing documentation).

Data is an essential component of ML. Yet, much of the focus in ML and NLP is on the specifics of the architecture. Model-centric approaches prioritise state-of-the-art performances, which can, at times, be marginal and reliant on practices such as fixing random seeds and selective reporting (Henderson et al., 2018; Lipton and Steinhardt, 2019). Crucially, model results do not necessarily lead to scientific understanding.

In more recent years, there has been a push for data-driven/data-centric practises for artificial intelligence and ML (Miranda, 2021; Whang et al., 2023)[59]. Sambasivan et al. (2021) note the importance of data, particularly in the domains of health and conservation. Anik and Bunt (2021) describes the perceived transparency of explainable data for AI systems. This is a particularly significant point in the context of language documentation and low-resource languages.

[58] Mishra et al. (2022) propose a data quality metric for benchmarks in NLP.
[59] See Datacentricai for the project which has reinvigorated these efforts.

Data-centric research focuses on strategies for building and maintaining datasets effectively and ensuring data quality [60]. The focus on data can inform the design of representative resources and corpora. We currently lack best practices and principled methods for efficiently building datasets for different tasks whilst reliably ensuring data quality. Further, the goal of linguistic documentation is an overarching archive that includes cultural and linguistic records that adequately represent a speech community (Himmelmann, 1998). The latter half of this thesis explores these ideas further.

[60] Such as the Data-centric AI workshop (NeurIPS 2021)

Chapter 2

Finite-state Approach to Modelling Morphology

This chapter was published as:

Muradoglu, S., Evans, N., & Suominen, H. (2020, July). 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.

Author contributions: S.M. and H.S. designed research; S.M, performed research; N.E. contributed linguistic data; S.M., H.S. and N.E. analysed data; and S.M. wrote the paper and revised it, based on critical comments by H.S. and N.E..

The first step in developing computational resources for morphology begins with creating a finite-state description. In doing so, it is possible to leverage existing linguistic knowledge by enforcing a rule-based system that does not rely on large quantities of data to achieve high levels of accuracy. This paper outlines the development of

a finite-state transducer for the Papuan language Nen[1]. Given the infrastructure afforded by FSTs, the model can be used for glossing and predictively generating forms. Glossing involves morphological segmentation and analysis. These analyses can later be used to build IGTs. Prediction or inflected form generation can be utilised as a mechanism to examine whether the rules that are used to describe the language hold in every environment.

For example, consider the transitive verb *wakaes* 'to see'. The diphthong *'ae'* in the non-dual form is shortened to *'a'* (e.g., ***yaka****tan* '3sgU:see:IPFV:ND-1sgA' 'I see him/her'). For dual forms, the full diphthong *ae* is present, followed by dual-marking *{-w-}* (Evans, 2016). For example, ***takaewt*** '3sgU:see:IPFV:ND-2|3nsgA' 'you two/they two saw him/her'. Based on this information for the '3sgU: NEUT.PRIM:1duA' grammatical features, the model predicts *yambaewanzm*. The full diphthong of the verb stem, followed by the dual-marking {-w-}, is recorded for imperfectives, but it is yet to be attested in other TAM values, such as the neutral and perfective categories. This kind of large-scale testing can aid in identifying areas that need further attention.

The transducer presented focuses on verbal morphology. The reasons for this are two-fold; first, the verb is the most complex word class in the Nen language, with a transitive verb taking up to 1,740 distinct forms[2]. Second, the structural properties of Nen verbs raise interesting choices for analysis. Nen verbs exhibit distributed exponence[3] where multiple exponents are required to resolve the intended morpho-syntactic category. As the name suggests, often, these exponents are distributed throughout the word.

In addition to computational resource building, this paper further explores a 'Chunking'

[1]The code for the architecture is available at https `://github.com/smuradoglu/nqn-morph`. The morphological rules are adapted from Evans (2016). The data set has not been made publicly available as a default. This is because the language community is still to be consulted at the time of publication of this thesis.

[2]The derivation of this figure is outlined on p.40 in chapter 1.

[3]See the section 1.1 for more detail.

[4] and a decomposition approach. This distinction is motivated by what would be considered gold standard practices for linguists and what is convenient from a computational point of view. The maximal decomposition approach follows a more classical linguistics approach whereby the minimal units are utilised to construct a word.

While this approach allows for greater generalisability, it introduces complexities by requiring pruning of certain disallowed combinations. For example, the dual thematic (*{-anz-}* for the neutral preterite (from table 1.5) cannot combine with the singular and plural actor numbers listed in table 1.6, it can only combine with the dual actor person numbers. Further, not all person/number combinations are available for each TAM category (noting that the TAM category is partly distinguishable by the prefix). This restriction needs to be carried through the thematic and the desinence. Things become even more convoluted with the ∅ morphemes in both the thematic and desinence. In particular, the sequence *'ng'* could be parsed as either *{-∅-ng}* or *{-ng-∅}*. If no restrictions were applied, the model would over-assign the zero morphemes. These options are not linguistically viable because the TAM features do not match. Defining rules to block certain combinations leads to more rules (Beesley and Karttunen, 2003).

This issue can be circumvented by adopting a less strict approach, here called 'Chunking'. 'Chunking' refers to grouping morphemes into prefixes and suffixes while the same amount of meta-information is retained. In this approach, the non-dual thematic would be combined with the singular and plural actor desinences. The disallowed non-dual thematic and dual desinence combination is not a possible pathway. So the sequence *ng* could be analysed 'IPFV.NPHD: 2|3du>du', 'IPFV.YPST: 2|3du>du', 'IPFV.RMPST: 2|3du>du', 'PFV.IMM:2plA', 'PFV.FUT:3plA' depending on whether the prefix belongs to the α, β or γ category and context.

Both models are built using the FST toolkit foma (Hulden, 2009). The resultant

[4] Inspired by the definition: *'unanalysed exemplars or chunks'* from Kelly et al. (2014).

architectures show differences in size and structural clarity. The size, states, arcs and paths are reported, with and without the flag diacritics eliminated. The numbers correspond to an FST with only one verb stem. This is to focus on the morphological system instead of stem alternations. The 'Chunking' model is under half the (computational) size of the full decomposed counterpart, but the decomposition displays a higher structural order. The arcs in the figures shown in the paper note the flag diacritic relations. With the distributed exponence characteristic of Nen verbs, flag diacritics are a necessary way to code the long-distance dependencies between morphemes. For example, consider the verb *y-apr-tam* (the verb *waprs* 'to make' inflected for '3sgU:IPFV:NPHD:1plA'), the prefix {y-} is marked with a P.alpha.on, which sets a value of α (i.e., the dummy prefixing series). For the 'Chunking' approach, the suffix {-tam} is marked with R.alpha, which notes that the α prefix is required. The decomposition approach breaks the suffix into two parts: {-ta-} and {-m}. The thematic is marked with two flags. First, the R.alpha to ensure the suitable prefix is used. Second, it is marked with a P.NPHD.ND to enforce the alignment of the correct morphemes between the thematic and desinence. This last flag is resolved at the desinence with a R.NPHD.ND.

Both models are evaluated on a hand-annotated verb corpus. Unsurprisingly, the overall accuracy is 80.3%. This consistency in performance is expected when both strategies are correctly implemented — as such, it also confirms the integrity of the architecture constructed. The choice of model depends on the primary concerns of the user. A more comprehensive evaluation remains a future task. Here, the reported accuracy pertains to assigning the correct MSD based on the inflected form. Another evaluation might be on the accuracy of synthesising inflected forms from lemma and MSD. Examining overgeneration in either application may reveal further distinctions between the approaches considered. In particular, given the intricate relationship between the thematic and desinence suffixes, some illegal analyses may be overlooked.

From a computational perspective, there is not much difference between the two other than the size of FST. In terms of implementation complexity, the 'Chunking' approach is more straightforward. This approach lists all available combinations. By contrast, the decomposition approach identifies the minimal units, but instead of listing all available combinations, it requires blocking unavailable combinations. One factor to consider is the size of the transducer produced. This can be an important point depending on the intended use. If the aim is to use the resultant FST in the field, having a lightweight model can be very important, given that in-field computation resources are often low. Combining the full lexicon, complexity and number of rules can quickly lead to a heavy system. If the user prefers structural granularity or a one-to-one mapping between the computational implementation and the linguistic grammar (perhaps for explainability), then the decomposition approach can be taken. One of the prime reasons to develop an FST morphological description is to gloss existing texts. In this case, there is no computational motivation for having a high-resolution description. Further, one of the drawbacks of FSTs is the expertise required in both finite-state methodology and the subject language.

Beemer et al. (2020) compares FST-based modelling with deep learning infrastructure for the SIGMORPHON shared task for morphological inflection. The paper focuses on developing over 25 finite-state grammars. Of the 25 language models submitted, 13 matched or surpassed the strongest neural-based submission. Several of these were relatively 'easy' — as they did not contain complex morphophonology, and the inflection patterns were regular. The study finds that it is not easy to outperform state-of-the-art neural network models.

In cases where the hand-written grammars outperform the neural models (e.g., Tagalog and Ingrian), a significant development effort was required by trained linguists to analyse and describe morphophonological patterns. For some of the languages in the task, only a small subset of morphosyntactic features were included. In these cases, although developing a highly accurate hand-written grammar was a

straightforward task, the success was diminished by the apparent ability of neural sequence-to-sequence models to also capture such morphological systems with high accuracy and little data. The time and expertise required to build FSTs is significantly higher than that of neural approaches. For languages that exhibit high morphophonological complexity and a variety of inflectional classes, Beemer et al. (2020) estimates hundreds of hours of development effort by a trained linguist to potentially surpass the performance of a current state-of-the-art seq2seq model. With these differences in mind, the next chapter details the implementation of deep learning-based approaches to model Nen morphology.

To Compress or not to Compress? A Finite-State Approach to Nen Verbal Morphology

Saliha Muradoğlu[1,2], Nicholas Evans[1,2], and Hanna Suominen[1,3,4]

[1]The Australian National University (ANU) / Canberra, ACT, Australia
[2]ARC Centre of Excellence for the Dynamics of Language (CoEDL) / Canberra, ACT, Australia
[3]Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO) / Canberra, ACT, Australia
[4]University of Turku / Turku, Finland
Firstname.Lastname@anu.edu.au

Abstract

This paper describes the development of a verbal morphological parser for an under-resourced Papuan language, Nen. Nen verbal morphology is particularly complex, with a transitive verb taking up to $1,740$ unique features. The structural properties exhibited by Nen verbs raises interesting choices for analysis. Here we compare two possible methods of analysis: 'Chunking' and decomposition. 'Chunking' refers to the concept of collating morphological segments into one, whereas the decomposition model follows a more classical linguistic approach. Both models are built using the Finite-State Transducer toolkit foma. The resultant architecture shows differences in size and structural clarity. While the 'Chunking' model is under half the size of the full decomposed counterpart, the decomposition displays higher structural order. In this paper, we describe the challenges encountered when modelling a language exhibiting distributed exponence and present the first morphological analyser for Nen, with an overall accuracy of 80.3%.

1 Introduction

With the advance of modern technology, collecting data for the task of language documentation has become easier, but methods for coping with the influx of data have become a pressing concern. One robust solution in the realm of morphology and phonology has been Finite State methods.

This paper focuses on the development of Finite-State architecture in aid of the glossing process for building resources for Nen. Nen is a under-resourced language of the Morehead-Maro language family of Southern New Guinea (Evans, 2015). It is spoken by approximately 300–350 people in the village of Bimadbn in the Western Province of Papua New Guinea. The resources developed here feed directly into the efforts of documentation and corpus building. This effort is globally shared amongst fieldworkers and descriptive linguistics across many languages, in response to the estimation for half of the world's languages to be extinct within the next century (Krauss, 1992). Aside from aiding the documentation process, the linguistic property of multiple exponence (ME) makes Nen an interesting case study for computational methods, as well as exasperating the already present data sparsity problem.

Though much of the recent work in Natural Language Processing (NLP) has centred around machine learning, it is still not quite feasible in low resource problem sets. Neural networks remove the need for incorporating detailed knowledge of the specific context by optimizing the mapping between input/output pairs. As a consequence a large amount of training data is required (Gorman and Sproat, 2016). In the low resource language setting, often linguistic insight can be exploited to help generate larger datasets, such as Finite-State methods being used to produce labelled data for training of neural networks (Moeller et al., 2018).

Finite-state Transducers (FSTs) are widely accepted as a standard way to computationally model the morphological structure of words in natural languages (Beesley and Karttunen, 2003; Koskenniemi, 1983). Prior works include FSTs for agglutinating languages such as Turkish, Tuvan, and Northern Haida (Çöltekin, 2014; Tyers et al., 2016; Lachler et al., 2018), and more recently so-called polysynthetic languages like Chukchi, Kunwinjku, Central Siberian Yupik, and Arapahoe (Andriyanets and Tyers, 2018; Lane and Bird, 2019; Chen and Schwartz, 2018; Kazeminejad et al., 2017).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 207–213
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

The novel contributions of this paper are twofold: First, we present a preliminary morphological analyser for verbs in Nen. In addition to resource building for the Nen language, this work outlines a computational approach for modelling the linguistic phenomenon of distributed exponence.

2 The Nen Language

With on-going documentation efforts, the Nen corpus is approximately $30,000$ words of natural speech, of which there are approximately $6,000$ verbs tokens (Muradoğlu, 2017). Over a third of these verb tokens ($2,379$ tokens) are varieties of the copula, which form a restricted paradigm of their own. Simply put, the amount of data is scarce. To add to this problem, Nen exhibits complex verbal morphology. In fact, verbs are morphologically the most complicated word-class in Nen (Evans, 2016, 2019). Despite this, they are often regular, allowing for generalisation of rules to analyse them. As outlined by Evans (2016), Nen verbs can be divided into two categories: prefixing and ambifixing verbs. Prefixing verbs mark the undergoer argument by prefix and ambifixing verbs employ both prefixes and suffixes to index person and number of up to two arguments. In this paper, we focus on the more complicated case of the ambifixing verb. The full prefix and suffixal paradigm can be found in Evans (2016) Table 23.3 (pg 548), Table 23.14 (pg 563) and Table 23.16 (pg 565).

The undergoer prefixes are divided into arbitrarily labelled series α, β, γ, which do not correspond to specific semantic values until they are unified with other TAM (Tense, Aspect, and Mood) markings on the verb (Evans, 2015). 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 verb *armbs* 'to climb'. When marked for direction the resultant forms are as follows: *n-armb-te* '(s)he is ascending (neutral)', *n-n-armb-te* '(s)he is coming up (towards speaker)', and *n-ng-armb-te* '(s)he is going up (away from speaker)'.

The middle prefixes simply mark the verb as a member of the middle verb type; essentially dynamic monovalent verbs. Prefix cells with more than one entry note possible allomorphy depending on the phonological environment within the verb. The suffixal system applies to both middle and transitive verb types.

Although it is convenient to segment verbs, into prefix, stem, and suffix, the Nen verbal system distributes information in a complicated way. The prefixes and suffixes are not independent values. Nen exhibits a particular kind of multiple exponence (ME), which requires prefixes and suffixes to be unified before inflectional values are known (Evans, 2016).

The possible combinatorial space for transitive and middle verbs is determined by summing the forms associated with each series (α, β, and γ) and the TAM suffixes they can co-occur with. The figure obtained is then multiplied by the possible undergoer prefixes (with only three available to the middle verbs). Lastly, this number is multiplied by three for each directional prefix available. This process yields a $1,740$ cell paradigm size for the transitive verbs.

2.1 Distributed Exponence

One of the prime motivations for choosing Nen as a case study is the phenomenon that gives rise to this combinatorial power: distributed exponence.

In linguistics, the notion of extended exponence was first introduced by Mathews (1974) and is now commonly referred to as multiple exponence (ME). Matthews defined ME as "a category if positively identified at all, would have exponents in each of two or more distinct positions" (Mathews, 1974). Distributed exponence is a kind of ME, which involves the use of more than one morphological segment to convey meaning. It requires all relevant morphs to yield a precise interpretation of the feature value in question (Carroll, 2016; Harris, 2017).

(1) N-n-and-armb-ta-ng
M:α-VEN-FUT.IMP-Nsg-ascend-Ndu:IPF-NSG.IPF.IMP

'You|they (>2) climb up later! (in the future, said to a group of people)'

In the example above, no one marker marks the plural person. The information of the agent being plural is distributed across the thematic (dual/non-dual) and the desinence (single/dual/plural). If a non-dual thematic is present than the desinence cannot have dual features, and so the only options are singular or plural. Further, this is an example of the future imperative in Nen. The future imperative category is marked by an additional prefix,

which also carries information about the agent. It carves up the person space in a different way to the thematic, and yet these values must be compatible. The other main feature value evident in this example is the prefix n- which serves as a dummy variable to reduce the valency of the verb, but it also yields information about the membership of the class α. Together with the desinence (and in this case the presence of the future imperative prefix), the TAM feature can be obtained.

3 Method

Several implementations of FSM compilers were available: XFST (Xerox Finite-State Transducer) (Beesley and Karttunen, 2003), foma (Hulden, 2009), and HFST (Helskini Finite-State Transducer) (Lindén et al., 2011), of which the latter two are open source. To develop a morphological analyser for Nen, we employed the foma Finite-State toolkit.

FSTs are an ideal tool for morphology, since they allow for both analysis and synthesis, meaning the user can both decompose a word and construct one, given the desired morphological features. Additionally, given the ongoing nature of language documentation, linguistic rules are constantly being added to, reviewed and revised. The incremental modularisation of FSTs allows for easy testing of set rules and addition of new rules.

FSTs are constructed in two parts: the first part deals with morphological rules and irregularities, as well as lexicon creation. The second component implements morphophonological rules.

3.1 Long Distance Dependencies (LDDs)

As with most languages, there are long-distance dependencies (LDD) that need to be resolved. This is even more true of Nen given its distributed nature. In FSTs, the transition from one state to another depends on the current state and the next input symbol. To transition to a state at time $t + 1$, the only thing considered is the state at time t (i.e., Markov assumption). In other words, there is no stack or other memory-like function that can be consulted.

One way of introducing memory is through Feature-setting and Feature-unification operations. These are practically implemented using flag diacritics (Hulden, 2011). Arcs with flag diacritics are like an epsilon transition but are conditional on the success or failure of the operation specified by the flag. In our setup, the operations used are P (positive) and R (require). This process is often repeated through the verb, where the unification of features is required.

3.2 Future Imperative

In addition to normal imperatives, Nen has future imperatives. This type of imperative specifies that an action should be carried out at some later point, and often at a different location (Evans, ms)

As seen in example 1, the TAM category of future imperative requires another prefix. Essentially at this point the FST has three options, {-and-} for non-singular, {-ang} for singular and {-∅-}. If the verb is not a future imperative than the {-∅-} pathway is taken. The future imperative is only possible if the prefix is of the α class.

The Nen language distinguishes between SG, DU, PL persons. For the decomposition model, there needs to be restrictions for the thematic, which splits this combinatorial space in a different way: Dual (DU) or Non-Dual (ND). A non-singular future imperative prefix cannot be used with a singular actor suffix.

This licensing of information can be done in several ways. For simplicity, the LDD is recalled in the shortest way possible. If this prefix is present then the system knows the series must be α, so instead of propagating the series restrictions to the end, we require the FUT.IMP (SG/NSG) feature to be unified.

3.3 Models

In building an FST for the Nen verb, the question of whether to 'Chunk' or decompose arose. By 'Chunking', we refer to the idea of combining morphological segments rather than decomposing to the minimal units (as briefly mentioned in Lachler et al. (2018)).

There are several motivations for this distinction. First, from a technical point of view, decomposing requires more rules to govern the combinations of even more segments. By having to block the possibilities of certain combinations (i.e., negative definition), this leads to more complex rules which need to be carefully considered and tested.

Secondly, this distinction neatly parallels with psycholinguistic theories dealing with processing of agglutinative or polysynthetic languages. The basic idea is that there is a dual mechanism for processing inflected words: lexical memory and morphological decomposition/grammatical rules (Hahne et al., 2006; Ullman, 2004).

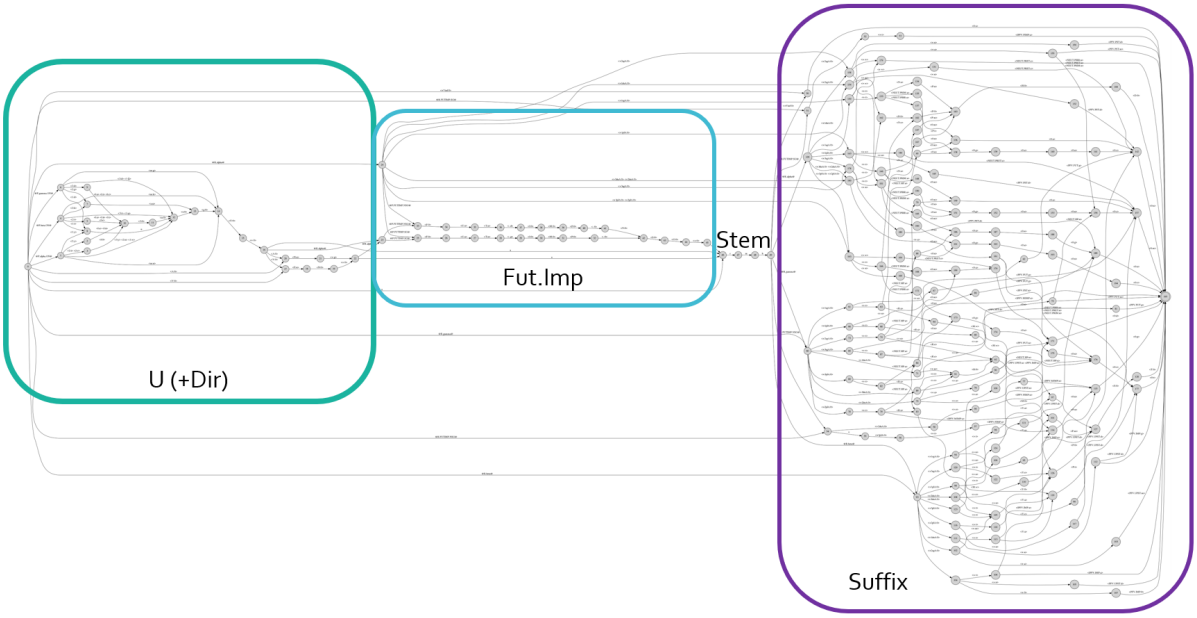


Figure 1: Overall FST architecture for 'Chunking' model. For larger view: 'Chunking'

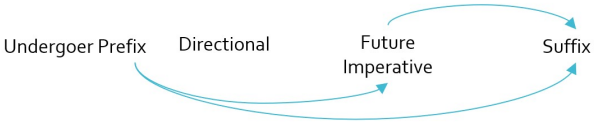


Figure 2: Information flow for the 'Chunking' model.

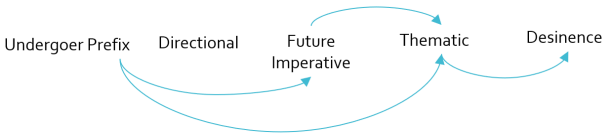


Figure 3: Information flow for decomposition model.

3.3.1 'Chunking' Model

As described above, 'Chunking' refers to the idea of combining morphological segments. In the case of Nen, this means treating the thematic and desinence as one rather than two separate segments. The thematic and desinence have the same hidden featural restrictions. That is to say, thematics of the same TAM feature can be unified with desinences of the same value. In this approach, the Undergoer prefix limits the possible allowed suffixes and forces certain TAM interpretations. Figure 2 depicts the LDD resolution for this model. We impose a prefix series restriction since the membership of the prefix (whether α, β, or γ) changes the interpretation of the suffix. It is a much more straightforward model compared with the decomposition model discussed next

3.3.2 Decomposition Model

The decomposition model follows the analysis of Evans (2016). It segments morphemes to their minimal meaningful units. This approach gives a more granular insight into the flow of information from one segment to the next. In fact, it is simply the uncompressed version of the 'Chunking' model. Decomposing into smaller units gives rise to more complex rules to constrain the FST to linguistically viable forms only. For example, Nen has {-∅-} and {-ng-} as possible thematic values, but it also has these same values in the desinence, so if no restrictions exist the system would over-assign the zero morphemes. The 'ng' suffix could be analysed as either {-∅-ng} or {-ng-∅}. Both these options are not linguistically viable because the TAM features do not match. In the decomposition model, we need to impose restrictions between all three: undergoer prefix, thematic and desinence (and the future imperative prefix). The simplest way to do this is to plan restrictions from undergoer prefix to thematic, and thematic into desinence (since they adhere to the same underlying paradigmatic structure) as seen in Figure 3. Instead of enforcing the dependency from the undergoer prefix, the range of the LDD or feature-unification is minimised. Since the future imperative and thematic already block the unsatisfactory feature-holding morphemes, the desinence only needs to be unified with the thematic morpheme.

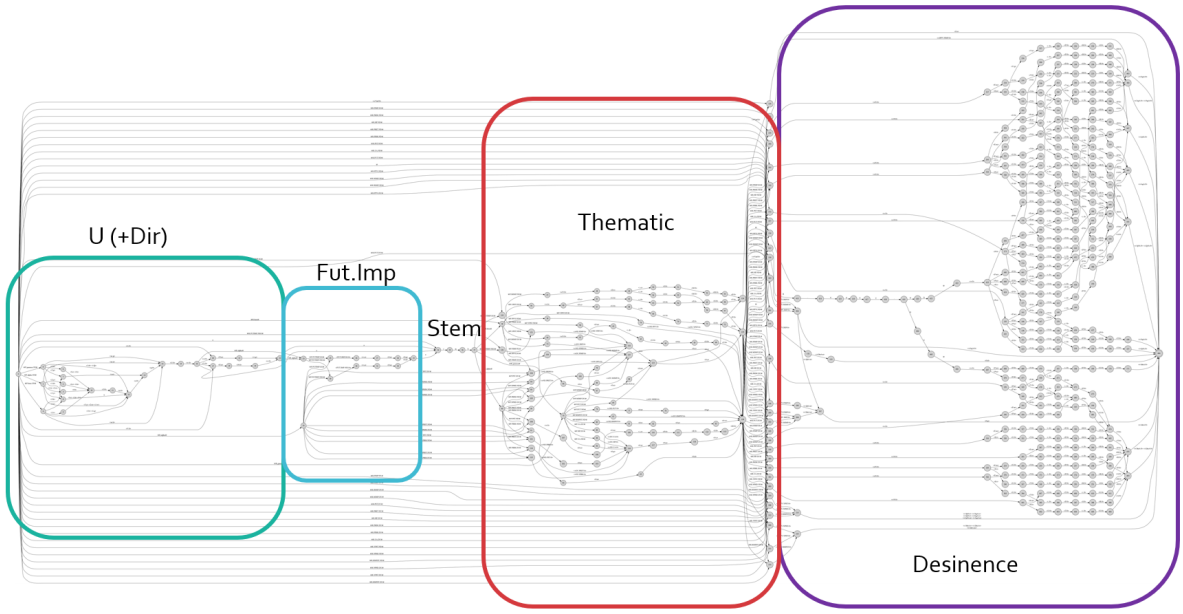


Figure 4: Overall FST architecture for decomposed model. For larger view: Decomposition model.

4 Results

The decomposition model showed a clearer level of organization than the 'Chunking' model (Figures 1 and 4, both with the flags included). Note that, one verb stem *armbs* 'to ascend' was used in both figures, for visibility of manifestation of morphological paradigm for one ambifixing verb. The particular stem was chosen because we had a full paradigm elicitation from members of the Nen community to confirm the existence of predicted forms. When comparing the specifications of both models, shown in Table 1, we could see that the decomposition was roughly double the 'Chunking' model in size, the number of states and arcs, and approximately 3.5 times more pathways.

These results questioned the benefit of decomposing further, apart from the obvious benefit of following the linguistic description. Given the added difficulty of implementing, if both yield comparable results, and the end goal is to have the highest possible accuracy of gloss than the choice of model should not matter.

4.1 Evaluation

We evaluated our FST models by comparing the glosses produced with those of a hand-annotated set (Muradoğlu, 2017). The hand-annotated corpus was derived from the Nen natural speech corpus. This included $1,680$ unique inflected forms (with the middle and transitive verbs making up approximately 58% of verbs observed) and 274 stems. Unsurprisingly, the hand-annotated corpus displays

Features	'Chunking'	Decomposition
Size	8.0kB (7.6kB)	13.7kB (15.2kB)
States	230 (197)	513 (470)
Arcs	385 (340)	709 (656)
Paths	$5,371$ $(26,288)$	$18,706$ $(811,069)$

Table 1: FST attributes for 'Chunking' and decomposition model with diacritic flags eliminated. Figures in brackets refer to the flag counterparts.

Zipfian properties, with the copula verb (and all of its inflections) being the most frequently occurring and making up 39% of the corpus. The coupla verb in Nen takes up to 40 unique forms which can be modelled perfectly.

During testing, we encountered an unexpected difference between the two proposed models.The definition of the imperfective basic non-dual thematic ({-taw-}|{-ta-}) required a morphophonological rule to drop the *a* or *aw* and attach the {-e} desinence for the 2|3sg actor. We addressed this problem in the *foma* file. This again, reiterates the notion of more rules required for further decomposition.

Both 'Chunking' and decomposition model showed an 80.3% accuracy (70.5% if only middle and transitive verbs are considered). The most common errors were attributable to spelling and/or morpholonological changes. For example, the inflected form *näramanda*, would only be recognised by the FST as *nrämnda* with the stem as *räm*. This

is because, exceptionally, the verb stem *(w)ärama-* 'to give' does not appear in full in the infinitive *räms*, whereas other verbs with benefactives (e.g. *wabens* 'to feed for') do include the prefix. The verb stem for give is built by adding benefactive {wä-} 'make' (thus 'giving' is literally 'doing for') to the root *räm* (infinitive *räms*) 'to do'.

Some of the unrecognised forms can be a result of variation in transcription. With ongoing efforts of documentation, transcription decisions evolve, resulting in a distribution of forms that represent the same thing. A typical example of this variation in the corpus is *wétélés|wetls* 'to tell/say/report', with the epenthetic vowels either being written orthographically or omitted. Typically these issues would be dealt with in the pre-processing stage however, some of these cases are harder to recognise than others, as is the case of handling naturalistic data.

5 Conclusion

This paper explores options for modeling the low-resource language Nen using finite-state transducers. Nen shows distributed exponence; multiple morphs can contribute to the specification of a particular feature value. This property motivates the comparison between a 'Chunking' model, which combines the thematic and desinence segment, to a decomposition model which handles the two separately at the cost of many more parameters. Both models achieve the same accuracy of 80.3%. The choice of model depends on the primary concern of the user. Assuming that either segmentation is linguistically possible, if the size of the transducer is of concern (as a result of the size of lexicon, complexity of rules or sheer number of rules) a 'Chunking' approach can be taken with no cost to accuracy. If the user, prefers structural granularity or a one-to-one mapping between the computational implementation and the linguistic grammar then the decomposition approach can be taken. Most often, the primary use of FST grammars are to provide morphological glosses, in this case there is no computational motivation for having a high resolution description.

Future work would entail analysing and implementing more detailed underlying morphonological rules, and investigating the cross-over from FSTs to neural models. One of the prime motivations for building an FST, in the era of neural networks is to generate enough labelled data, in the appropriate format to enable testing across architectures. Additionally, the process of building an FST proves to be a great way to examine the validity of the linguistic analyses.

Acknowledgments

We are grateful for the mentoring scheme provided by the ACL student research workshop. In particular, we would like to thank Greg Durrett and Richard Sproat for their constructive feedback during the mentoring phase.

References

Vasilisa Andriyanets and Francis Tyers. 2018. A prototype finite-state morphological analyser for Chukchi. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 31–40, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite-state morphology: Xerox tools and techniques*. CSLI, Stanford.

Matthew J. Carroll. 2016. *The Ngkolmpu Language*. Ph.D. thesis, The Australian National University.

Emily Chen and Lane Schwartz. 2018. A morphological analyzer for St. Lawrence Island/Central Siberian Yupik. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Çagri Çöltekin. 2014. A set of open source tools for turkish natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1079–1086.

Nicholas Evans. 2015. Valency in Nen. In Andrej Malchukov, Martin Haspelmath, Bernard Comrie and Iren Hartmann, editors, *Valency classes: A comparative handbook*, pages 1069–1116. Berlin: Mouton de Gruyter.

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. 2019. Waiting for the word: distributed deponency and the semantic interpretation of number in the Nen verb. In Andrew Hippisley Matthew Baerman, Oliver Bond, editor, *Morphological perspectives*, pages 100–123. Edinburgh: Edinburgh University Press.

Nicholas Evans. ms. Grammar of Nen.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Anja Hahne, Jutta L. Mueller, and Harald Clahsen. 2006. Morphological processing in a second language: Behavioral and event-related brain potential evidence for storage and decomposition. *Journal of Cognitive Neuroscience*, 18(1):121–134.

Alice C Harris. 2017. *Multiple exponence*. Oxford University Press.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Mans Hulden. 2011. Morphological analysis tutorial:a self-contained tutorial for building morphological analyzers.

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18.

Kimmo Koskenniemi. 1983. *Two-level morphology*. Ph.D. thesis, Ph. D. thesis, University of Helsinki.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur Moshagen, and Antti Arppe. 2018. Modeling northern haida verb morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

William Lane and Steven Bird. 2019. Towards a robust morphological analyzer for kunwinjku. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 1–9.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg. 2011. Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Peter H Mathews. 1974. *Morphology: an introduction to the theory of word-structure*. Cambridge, England: Cambridge University Press.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer. pages 12–20.

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.

Francis Tyers, Aziyana Bayyr-ool, Aelita Salchak, and Jonathan Washington. 2016. A finite-state morphological analyser for tuvan. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2562–2567.

Michael T. Ullman. 2004. Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1-2):231–270.

Chapter 3

Using Neural Networks to Model Morphology

This chapter was published as:

Muradoglu, S., Evans, N., & Vylomova, E. 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.

Author contributions: S.M. and E.V. designed research; S.M. performed research; N.E. and S.M. contributed linguistic data; S.M. and E.V. analysed data; and S.M. wrote the paper and revised it, based on, critical comments by N.E. and E.V..

The time and expertise required to build FSTs are significantly higher than that of neural approaches. For languages that exhibit high morphophonological complexity and a variety of inflectional classes, (Beemer et al., 2020) estimates hundreds of hours of development effort by a trained linguist to construct an FST to *potentially* surpass the performance of a current state-of-the-art sequence-to-sequence model.

With these differences in mind, this chapter explores the application of deep learning approaches to analysing Nen verbal morphology. Two high-performing submissions

from the SIGMORPHON–CoNLL 2017 Shared Task (Cotterell et al., 2017) are utilised to examine the feasibility of this extension. This choice is based on both general performance and, more pertinently, the success shown in the low-resource (100 training samples of inflection) setting.

From an NLP perspective, diverse languages and their corresponding linguistic features are vital for testing the generalising capacity of models. To that end, the robustness of the two selected models (Aharoni and Goldberg, 2017 and Makarov and Clematide, 2018b) is tested by training both on Nen verbal morphology.

In this paper, the first-ever neural network-based analysis of Nen is presented, the first representation of the Yam language family and, to the best of our knowledge, of a Papuan language. Nen provides an interesting case study for neural approaches as it exhibits non-monotonic morphological mapping: distributed exponence (see 1.1 for a detailed example).

Further, this study addresses several questions: (1) how model performance varies across both training data size and sampling methods, (2) how training data composition (with respect to verb type) affects prediction accuracy, and (3) can these models infer properties of the language which are not annotated in the data.

Model performance is compared across both data size and sampling methods. The data sizes considered range from low-resource (100 samples), medium (1,000 samples), ALL (everything that is available from the corpus after test/dev split) and finally, high-resource (10,000 samples, obtained by data hallucination).

With the rare opportunity afforded by the direct distillation of glossed data from the corpus, comparing a more naturalistic 'Zipfian sampling' to the more widespread approach in NLP – random sampling is possible. A detailed taxonomy of the errors produced by the model is also given. 'Zipfian sampling' describes a sampling strategy that considers observation frequency. For example, the 100 most frequently occurring inflected forms are sampled for the training data in the low-resource setting. Note

that frequency is only used as a sorting metric and is otherwise not considered in the UniMorph-like (i.e., type-based) dataset created for Nen. Given the type-based dataset, each inflection is treated as equally probable. So, a random sample is unlikely to follow the same Zipfian distribution observed across the texts in the corpus. As Cotterell et al. (2018) notes, '*this is a realistic setting since supervised training is usually employed to generalise from frequent words that appear in annotated resources to less frequent words that do not*.'

Unsurprisingly, the best performance for both models is observed in the high-resource setting. Training data collated by random sampling yields slightly higher accuracies than their Zipfian counterparts. An extensive analysis of the types of errors generated by each system is provided. The most common error type was found to be allomorphy errors, a misapplication of morphophonological rules, or feature category mappings.

The errors are reported in a hierarchical manner as follows: Target > Stem > Allomorphy > Free Variation. This is in acknowledgement of some errors being worse than others. For example, if a predicted form has a stem error, this means that the generated stem is either a re-mapping of a seen but irrelevant stem or a made-up, nonce stem. This error type supersedes allomorphy as a stem error is arguably a more grievous faux pas than an allomorph error since the underlying lemma has changed. Further, it becomes a question of 'what does the correct allomorph for this new word look like?'. To categorise this error as an allomorph error, the evaluator would need to perform a wug test-like task (Berko, 1958)[1]. While it is possible to provide a prediction, language expertise is required to assess whether it is correct. A more comprehensive error distribution whereby all errors are reported remains a task for future work.

The minor differences in performance observed across the two sampling strategies may be attributed to training set composition differences. In the Zipfian case, the prefixing verb types are over-represented as they are more frequent in natural speech.

[1]When the inflected form is not observed in the corpus, this holds for mismatched stems as well.

This observation prompted a more detailed investigation of training data composition. Given that the inflectional paradigm size varies significantly across the verb types, it is conceivable that having more examples for the smallest paradigm (i.e., the prefixing verbs) might negatively affect model generalisation. In particular, since the only relevant part of the prefixing verb paradigm for the ambifixing verbs is the prefixal paradigm (see 1.1 for discussion). This experiment examines the transferability of learning about the inflectional paradigm from one verb to another.

A new subcategory of error, *'free variation'*[2] is introduced. In linguistics, free variation describes the case when two (or more) forms occur in the same environment without a change in meaning and without being considered incorrect by native speakers. This type is only an 'error' from the model's perspective, as it compares the predicted forms with a gold set. This type of evaluation assumes a convergent language standard. From a linguistic point of view, these variations of language are not considered errors as they are deemed grammatical by native speakers. Additionally, language variation can be conditioned by a whole plethora of factors, and these factors are typically the subject of study in sociolinguistics. When these 'errors' occur, either inflected form is typically in the top two most likely forms of the beam search — the probabilities are more likely an indication of form frequency observed in the corpus than any linguistic signal suggesting a base form. This approach can be extended to sociolinguistic analysis, although a larger corpus size would be required to make any statistically significant conclusions.

Unsurprisingly, when the training set contains only one type of verb, it performs best for the type of verb seen in the training data. The partitioned training sets show no particular evidence of a principal parts-like modular learning. This might be due to the small sample size, the limitation of the MSDs (explored further in chapter 4) or that the neural network generalisation does not follow the plausible linguistic

[2]The example given is a little misleading in the paper. It describes non-standard orthography where the epenthetic vowel is written– a relic from older transcriptions. Perhaps a better example would be *ym* and *ymn* 'NPST:2|3pl' from the copula paradigm (shown in table 1.3).

trajectory.

The third question explored in this paper is whether the models can infer properties of the language that are not annotated in the data. The test case is designed around the second- and third-person singular actor past perfective ('PFV.PST:2|3sgA') cells of the paradigm. While Nen verbs typically exhibit syncretism in the second and third cells of the paradigm, this is the one place where this pattern is broken. During training, the model is exposed to examples of the third-singular past perfective ('PFV.PST:3sgA') MSDs to test whether the model would overgeneralise and predict a syncretism that does not exist. At test time, the model is asked to predict the previously unseen second-singular past perfective ('PFV.PST:2sgA') form.

The results show that the model overwhelmingly predicts syncretism for the second-singular past perfective — much like a linguist or a human learner might guess. These results suggest that neural-based approaches learn structural information beyond a sequence of characters correlating to a particular MSD.

With structural inference across the paradigm shown possible, it rekindles the idea of minimum description length but with a focus on data rather than model architecture. In other words, how much data is minimally needed to describe the full system, and what does it look like? Instead of focusing on the model architecture, this approach embraces the data-driven nature of neural networks (in the supervised machine learning case). The next chapter navigates these ideas and questions by comparing a linguistic corpus-building approach with a dataset built specifically for NLP tasks such as morphological inflection.

Modelling Verbal Morphology in Nen

Saliha Muradoğlu[ΩΦ] **Nicholas Evans**[ΩΦ] **Ekaterina Vylomova**[μ]
[Ω]The Australian National University (ANU) [μ]The University of Melbourne
[Φ]ARC Centre of Excellence for the Dynamics of Language (CoEDL)
saliha.muradgolu@anu.edu.au, nicholas.evans@anu.edu.au,
ekaterina.vylomova@unimelb.edu.au

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.

1 Introduction

A long-standing research direction in NLP targets the development of robust language technology applicable across the wide variety of the world's languages. Unfortunately, the vast majority of machine learning models are being developed for a small fraction of nearly 7,000 languages in the world, such as English, German, French, or Chinese. With introduction of highly multilingual corpora such as UniversalDependencies (Nivre et al., 2016) and UniMorph (Sylak-Glassman et al., 2015; Kirov et al., 2018) the situation started to change. For instance, SIGMORPHON organized a number of shared tasks on morphological reinflection starting from 10 languages in 2016 (Cotterell et al., 2016) and up to 90 languages in 2020 (Vylomova et al., 2020). In 2020, languages were sampled from various typologically diverse families: Indo-European, Oto-Manguean, Tungusic, Turkic, Niger-Congo, Bantu, and others. Still, just one language, namely, Murrinh-patha, an Australian Aboriginal language (Mansfield, 2019), represented the whole linguistic variety of the Oceania region. In this paper, we aim at filling the gap by exploring Nen, a Papuan language spoken by approximately 400 people in Papua New Guinea. Nen is known for its rich verbal morphology, with a transitive verb inflecting for up to $1,740$ feature combinations. Distributed exponence, the phenomenon which gives rise to this large paradigm size, provides insight into modelling complex mappings between surface forms and feature bundles.

We conduct a series of experiments on morphological reinflection task recently introduced under the umbrella of SIGMORPHON (Cotterell et al., 2016, 2018). We train several state-of-the-art machine learning models for verbal inflection in Nen and provide an extensive error analysis. We investigate the relationship between the distribution of verb type (inflection classes) in the data and performance. Finally, we show that the system learns properties of the data that are not explicitly given, but may be inferred.

The rest of the paper is organized as follows: In Section 2, we give a brief overview of related work. Section 3 provides an overview of Nen verbal morphology, Section 4, details our methodology, and Section 5 presents our results. Finally, Section 6 concludes the paper.

2 Related Work

Muradoglu et al. (2020) is the only reported work on the computational modelling of the Nen language. Similar to this study, the main focus is on modelling Nen verbal morphology, but using finite-state architecture instead. The accuracy achieved by the FST system is 80.3% obtained across the corpus, with approximately 10% of the accuracy attributable to the modelling of prefixing verbs (the regularity of copula verbs boosts the accuracy from 70.5%). The accuracies reported are not directly

comparable with those presented here due to the different data splits, and increased amount of data.

In our error analysis, we follow the error taxonomy proposed by Gorman et al. (2019) upon a detailed analysis of typical errors produced by morphologically reinflection systems. A similar study was conducted for Tibetan (Di et al., 2019).

3 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). Most inhabitants are multilingual, typically speaking several of the neighbouring languages.

The subject of this paper – verbs – are the most complicated word-class in Nen (Evans, 2015, 2019b). They are demarcated into three separate categories: prefixing, middle, and ambifixing verbs. The latter two are mostly regular in terms of morphophonological rules. In the remainder of this section, we elaborate on these characteristics, to give the reader enough background to follow the discussion in subsequent sections.

3.1 Verbal morphology

We begin our description from the maximal case – transitive ***ambifixing*** verbs. Examples of this verb type include *yis* 'to plant' and *waprs* 'to do' These verbs allow for full prefixing and suffixing possibilities. Evans (2016) provides the canonical paradigms for the undergoer prefixes, thematics and desinences. Suffix combinations are constructed by concatenating the corresponding thematic and the desinence. Between the undergoer prefix and verb stem is a directional prefix slot, available for all verb types. This slot is occupied by {*-n-*}[1] to convey a 'towards', {*-ng-*} for 'away' or left empty to convey a directionally neutral semantic.

Middle verbs such as *owabs* 'to speak' or *angs̄* 'to return', are also ambifixing, but the prefixal slot is restricted to {*n-*} (α–series), {*k-*} (β–series), {*g-*} (γ–series). These prefixes are person and number invariant, and mark the verb as being a dynamic monovalent verb. The prefix set is divided through the use of arbitrarily labels: α, β, and γ. These dummy indices do not carry specific semantic values until they are unified with other TAM (Tense, Aspect, and Mood) markings on the verb (Evans, 2015).

Prefixing verbs have separate closed paradigms, tailored to the subtype. Prefixing verbs are mostly distinguished through semantics; positional verbs such as *kmangr* 'to be lying down', the verb 'to own/have' *awans*, the verb 'to walk' *tan* and the copula verb *m* with its directional variants (be hither (i.e. come) or be thither (go)).

Inflectional prefixes for these verbs, mostly resemble the process with ambifixing verbs, yet the suffixes are limited. Of the 50 or so prefixing verbs, the vast majority are positional (Evans, 2020). An additional distinguishing feature of prefixing verbs, is the lack of infinitives. Both ambifixing and middle verbs form infinitives through suffixing *-s* to the verb stem. For the purposes of this study, we have listed the prefixing verb lemmas as the verb stem.

Methodologically, it is more convenient to segment a word as a classical bijective mapping between form to meaning. However, the Nen verbal system distributes information in a more complicated way. The prefixes (undergoer and future imperative) and suffixes (thematic and desinence) are not independent values. Nen verbal morphology is characterised by *distributed exponence (DE)*; "morphosyntactic feature values can only be determined after unification of multiple structural positions" (Carroll, 2016).

There are two consequences for morphological parsing:

a) Provisional unspecified values occur regularly, whether

   (i) These involve partial specification that will be filled in later in the word-parse, such as the left-edge prefix {*yaw-*} (1st person non-singular undergoer), which will only be made more precise in its number value (dual, or plural) when the thematic is encountered after the verb stem: thus *yaw-aka-t-an* 'I see them[2] (more them two)', where the 'non-dual' marker {*-ta-*} eliminates the dual (them two) but *yaw-akae-w-n* 'I see them

[1] We follow linguistic convention with '{}' denoting morphemes, and examples are italicised.

[2] Can also mean 'I see you (more than two)', resolved by combining with an appropriate free pronoun, *bm* 'you (absolutive)', but for present purposes we ignore this further complication.

(two)', where the 'dual thematic' {*-w-*} eliminates the plural (them more than two) reading.

(ii) These involve semantically-unspecified prefix series which only acquire meaning when they are combined with suffixes at the other end of the word: thus {*yaw-*}, in the above example, belongs to the α-series which, if it combines with the 'basic imperfective', will be given a (broadly) non-past reading, but when it combines with the 'past perfective' it will be given a past reading and when it combines with a 'projected imperative' it will be given a future meaning; a β-series form like {*taw-*}, by contrast, will have a 'yesterday past' interpretation when combining with the 'basic imperfective' suffixes but when combining with imperatives it will have a 'now/immediate command' meaning

b) More problematically, prefixes that normally have one reading (such as the yaw-example just discussed, which normally marks second/third person non-singular objects) sometimes have to be given a different meaning (e.g. large plural intransitive subjects) if further parsing to the right encounters a 'middle' rather than a 'transitive dynamic' stem (Evans 2017, 2019).

In principle that this means left-to-right morphological parsing is sometimes non-monotonic (particularly in the case of (b)), so that semantic values, as parsing proceeds, need to be sometimes held as provisionally unspecified, sometimes as partially specified, and sometimes as specified but subject to later override.

3.2 Distributed Exponence

One of the primary motivations for choosing Nen as a case study is the phenomenon that gives rise to this combinatorial power: distributed exponence. Essentially distributed exponence is a morphological phenomenon that gives rise to some types of non-monotonicity.

In linguistics, the notion of extended exponence was first introduced by Mathews (1974) and is now commonly referred to as multiple exponence (ME). Matthews defined ME as a category that would have exponents in two or more distinct positions. Distributed exponence is a kind of ME, which involves the use of more then one morphological segment to convey meaning. It requires all relevant morphs to yield a precise interpretation of the feature value in question (Carroll, 2016; Harris, 2017).

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

In the example above, no one marker marks the singular person. The information of the agent being singular is distributed across the thematic (dual/non-dual) and the desinence (single/dual/plural). If a non-dual thematic is present than the desinence cannot have dual features; the only options are singular or plural. Another morpheme present in this example is the prefix -ng- which marks the verb with the directional *thither*. The prefix n- marks this verb as a middle verb; it reduces the valency of the verb and yields information about the membership of the class α. Together with the prefix, thematic and desinence, the TAM feature can be obtained.

4 Methodology

4.1 Morphological reinflection task

Morphological inflection is a task of predicting a target word form from a corresponding word lemma and a set of morphosyntactic features (specifying the target slot, e.g. its part of speech (POS), tense, number, gender). For instance, a system is provided with a lemma "to sing" and a set of tags "Verb; Past" and needs to generate "sang". Morphological *reinflection* is a variation of the task when a lemma form is replaced with some other form and (optionally) its tags. The task has been traditionally solved with finite-state transducers, either hand-engineered (Koskenniemi, 1983; Kaplan and Kay, 1994) or trainable models that rely on both expert knowledge and data (Mohri, 1997; Eisner, 2002). In 2016 SIGMORPHON started a series of shared tasks on morphological reinflection, and neural models demonstrated superior performance when compared to finite-state or rule-based approaches, especially in high-resource languages (Cotterell et al., 2016; Vylomova et al., 2020).

[3] Example adapted from (Evans, 2020)

4.2 Data

The data used in this study comes from a Nen verb corpus (approximately $6,000$ verb samples representing $2,231$ unique inflected forms) created by Muradoğlu (2017). This dataset is a distilled subset from the approximately 8-hour natural speech corpus for the Nen language. As such it entails a frequency sorted list of all the verb forms occurring.

The training data is a set of triples comprising a lemma, morphosyntactic features, and an inflected form (i.e. we will only focus on morphological *inflection*).

Sampling Following the methodology in Cotterell et al. (2018) we split the data into training, development, and test sets. Training splits were created by sampling without replacement for three set sizes: all (*ALL*), medium (*MR*), and low (*LR*).

In virtue of coming from a natural corpus, the list of verb forms we use is Zipfian. This study does not distinguish between the feature bundles and only considered surface (inflected) forms. To facilitate the nature of our study, we uniformly distribute frequency across each syncretic cell.

For the *ALL* training set we start by sampling the first $1,931$ forms, in accordance with the Zipfian ranking across the corpus. In other words, we sample the $1,931$ most frequent verb forms. We randomly shuffle the remaining 300 forms into a 200 form test, and 100 form development (dev) sets. The test and dev sets remain the same through this experiment. Zipfian sampling is considered more realistic in this case, as it mimics the stimulus a language learner encounters. The dev and test set are randomly shuffled since supervised methods usually generalise from frequently encountered words.

For the *LR* and *MR* settings we take the first 100 and $1,000$ forms from the *ALL* training set, respectively. In addition, we create a high-resource (*HR*) set by supplementing the *ALL* set with synthetic forms, the final set contains $10,000$ forms. In order to generate synthetic samples, we use data hallucination technique proposed in Anastasopoulos and Neubig (2019). Note that the low-resource (*LR*) training set is a subset of the medium-resource (*MR*), which is supersetted by the *ALL* (and by extension the high-resource (*HR*) data set).

Finally, we contrast Zipfian sampling, when forms are sampled based on their frequency, to random sampling. Both sets (*LR* and *MR*) for the random sampling are created in a similar manner to Zipfian sampling, except frequency is not considered. Note that due to initial data size constraints, the *ALL* (and, therefore, *HR*) data sets for *both* the Zipfian and random sampling are the same. [4]

4.3 Experiments

In the current study we conducted three experiments to address our research questions.

4.3.1 Experiment 1: Testing across various data sizes and sampling methods

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

We evaluate modelling accuracies across four different training sizes, which is further contrasted across sampling type. Our experimental setup mirrors those of the SIGMORPHON reinflection tasks (Cotterell et al., 2016, 2017, 2018; Vylomova et al., 2020): given an input lemma and a set of feature tags, models generate inflected forms. The final accuracy is computed as the percentage of matches between the gold and predicted forms.

4.3.2 Experiment 2: Testing compositionality of training data

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

We test the effects of the verb type composition (i.e. how much of each verb type there is) in the training set. This study consists of seven (arising from all combinations of the three verb types) training data sets obtained through the sampling methods outlined above. We compare training sets of ambifixing verbs only, prefixing verbs only, middle verbs only, a two-way combination of each verb class: ambifixing and prefixing verbs, ambifixing and middle verbs, and prefixing and middle verbs and, finally an equal distribution of all three verb types, as listed in Table 4. Each set contains 386 forms (instances), stipulated by the amount of prefixing verbs available. The test and development set are 100 forms each, and is made up of 34 ambifixing, 33 middle and 33 prefixing verbs [5]

[4] Since the test and dev set are the same for both sampling methods, and are generated from the **remaining** 300 tokens (i.e. the least frequent items), it renders the random sampling of the *ALL* (and thus *HR*) the same.

[5] Uniform distribution is unlikely in natural language, in fact, Muradoğlu (2017) shows that the distribution is skewed

4.3.3 Experiment 3: Testing syncretism

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

In Nen, the second and third person feature bundles often correspond to the same surface form across the available TAM categories (i.e. are syncretic). We test the likelihood of both models predicting the *unseen* second person singular for the past perfective TAM category as syncretic with the *seen* third-person singular variant. This is the one instance across the Nen verbal paradigm where this syncretism does not hold. In essence, we examine linguistic patterns that may be inferred from an annotated dataset.

The main focus here, is to categorise the type of prediction rather than the overall accuracy, as such training and development sets are identical to those generated for the *ALL* setting in the first experiment. The test set is comprised of 100 inflections of the past perfective second singular tags, most of these have been gathered from the Nen dictionary (Evans, 2019a).

4.4 Models

For our experiments, we will utilise two models that have shown superior performance in SIGMORPHON–CoNLL 2017 Shared Task on morphological reinflection in low- and medium-resource settings (Cotterell et al., 2017). Both of them are essentially neural sequence-to-sequence models implemented in Dynet (Neubig et al., 2017). In addition, we also compare the results with a simple non-neural baseline used in 2017–2018 tasks on morphological reinflection (Cotterell et al., 2017, 2018).

Hard Monotonic Attention (Aharoni and Goldberg, 2017) An external aligner (Sudoh et al., 2013) first produces transformation operations between an input (lemma) and a target (inflected form) character sequences. The alignment operations (steps) are then fed into a neural encoder–decoder model. The network, therefore, is trained to mimic the transformation steps, and at inference time it predicts the actions based on the input (lemma) sequence. Unlike soft attention models, this model attends to a single input state at each step and either writes a symbol to the output sequence or advances its pointer to the next state. Hard attention models demonstrate superior performance in languages that employ suffixing morphology with stem changes.

Neural Transition-based (Makarov and Clematide, 2018) The model is essentially derived from Aharoni and Goldberg (2017) by enriching it with explicit insertion, deletion or, alternatively, copy mechanisms. The copy mechanism led to significant accuracy gains in low-resource settings. Following Rastogi et al. (2016), the model can be seen as a neural parameterization of a weighted finite-state machine.

Non-neural Baseline (Cotterell et al., 2017, 2018) The non-neural system first aligns lemma and inflected form strings using Levenstein distance (Levenstein, 1966) and then extracts prefix- and suffix-based transformation rules.

4.5 Settings

The hyperparameters of the models are set to the values reported in the corresponding papers as per Table 1.

Hyperparameters	A&G	M&C
Input dim	100	100
Hidden dim	100	100
Epochs	100	50
Layer	2	1

Table 1: Hyperparameters for both A&G (2017) and M&C (2018) models.

5 Results

Table 2 shows the accuracies achieved for each system for each training set size and sampling type from Experiment 1. For all setups the M&C model performed best with random sampling (where applicable). As expected the high-resource setting performs best overall. The random sampling yields slightly higher accuracies than the Zipfian counterpart, this is likely due to the fact that prefixing verbs, particularly the copula and its 40 distinct forms occupy a majority of the top 100 positions in the Zipfian distribution. Thus when random sampling is utilized the training set includes more examples of ambifixing verbs.

to favour a higher number of ambifixing verbs in terms of the number of inflected forms.

	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

Table 2: Data set, model and sampling accuracies. ALL is a total of 1,931 verbs, HR is 10,000, MR is 1,000 and LR is 100 samples for the training set.

	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	174†	22	0
Total	122	98	198	78	70	168	143	116	198	199	194	198

Table 3: Absolute number of errors on the test set (200 instances) made by each system trained in ALL, HR, MR and LR setting.*contains 5 looping errors,† 17 looping errors.

5.1 Error Analysis

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

We have taken a hierarchical approach to our error classification; whereby if more than one error is present, the category higher up is reported. For example, if a predicted form exhibits both target and allomorphy errors (error types are described in the following subsections), then only the target error is reported. The motivation for this lies in the nature of the error; free variation is technically not even an error. By contrast, misapplication of a morphophonological rule does indeed yield an incorrect form. Additionally, we have marked Target errors higher up as the system cannot be expected to correctly predict a form if the gold standard is incorrect. The hierarchy is as follows: Target>Stem>Allomorphy>Free Variation. 3 Table 3 summarises the types of errors across the different training sizes for each model. Overall, for both systems allomorphy errors remain relatively unimproved between the *ALL* and *HR* setting, but show a leap of reduction from the LR to MR conditions. Free variation errors are more prevalent in the *ALL* setting. This is probably a consequence of seeing more of the golden data and thus observing more of the systematic variations. This also explains why these errors reduce in number for the *HR* setting. The target errors are consistent across each experiment, as these are systematic issues with the gold data. Interestingly, stem errors reduce in the HR setting. This is despite the use of hallucinated data.

5.1.1 Allomorphy

This category consists of errors which are characterised by a misapplication of morphophonological rules, or feature category mappings. Frequent errors include the absence of vowel harmony or place assimilation rules, and incorrect mapping of feature bundles to surface forms. Most errors are of this category.

Vowel harmony. The Nen language exhibits vowel harmony. Consider the form *ynḡite* generated by one of the models, in a canonical sense the inflection is correct, but the presence of the high front vowel *i* requires the general *e* to harmonize to become *ynḡiti*.

Morphophonological Rules. When combining *r* final stems with *t* phonemes (which occurs in inflections via the non-dual thematics or certain desinences with ∅ thematics), the resultant sound is *n* (Evans, 2016). The M&C systems predicts that the stem *tar* inflected for the non-prehodiernal, first person actor and third person undergoer as *ytaretan*. Presumably, the break down is *y-tar-e-ta-n*. Interestingly, it inserts an *e* between the *r* and *t*, rather than concatenates the stem with the {-ta-n}

suffix. The correct form is *ytanan*.

Misapplication of category. These errors are rather straightforward: they are a misapplication of inflection rule and result in an incorrect cell of a paradigm. For example, *ynrenzan* is generated instead of *ynrenzng*. Technically, the form generated is correct, but it should correspond to the past perfective, first-person singular acting on dual actor suffix. Instead, it is mapped to the imperfective non-prehodiernal, third singular acting on dual actor suffix.

Future Imperatives. In all settings, across all systems tested, the future imperative was incorrectly predicted. Much like the β and γ counterparts, the system generated this TAM category by simply choosing an α prefix and suffixing {-ta}. Both A&G and M&C systems produce *yngita* instead of *yngangwita*. This formulation is correct for the β and γ series producing the imperfective imperative and mediated imperative, respectively. However, the future imperative has a special prefix which prefixes after the undergoer and directional prefix. It signifies the future imperative TAM category and marks the agent as either singular or non-singular.

Prefixing Verbs. Given the sparsity of examples for prefixing verbs and in particular their subtypes, a common occurrence across the data sizes is for the prefixing verb predictions to be inflected with the wrong features. For example, when the verb *m* 'to be' is inflected for the andative, 3PL+ undergoer and imperfective non-prehodiernal TAM the correct inflected form would be *yenewelmän*, instead the system gives *ynm* which it correctly identifies as a related form, but it does not have the correct inflectional features.

5.1.2 Free variation

Free variation errors occur when more than one acceptable inflected form exists; this is particularly true of the data set used in this study. The corpus used here has been distilled from a natural speech corpus, that has been transcribed. In addition to spelling variation - that arose as the orthographic decisions changed with ongoing documentation, the corpus also exhibits inter-speaker variation. An example includes: *yérniwi* as the predicted form and *yrniwi* as the gold standard. In Nen orthography, epenthetic vowels are not written in as their locations can be predicted (Evans and Miller, 2016). Older transcriptions wrote these vowels in with the *é*.

5.1.3 Target

These errors are characterised by incorrect feature tags in the gold standard data. One such example is as follows: the model predicts the form to be *nnganztat* and the gold standard is given as *ynganztat*, the feature tag, however, includes [M][6] and not [3SGU]. In such cases, based on the feature bundle, the system generated form is correct. This particular mismatch of middle and transitive verbs is the main source of this kind of error; it arises from the fact that a single verb may have a middle and transitive verb variant. This distinction can be difficult to decipher, and on some occasions, it can even be a result of speaker error.

5.1.4 Stem

This category denotes either a generated stem or a re-mapping of a seen but irrelevant stem, to the inflected form. These errors have linguistically viable morphemes attached, but we have not evaluated the accuracies of the mapping between feature and form for the morphemes.

Remapping

One such example is A&G model generating *ygmtandn* for the stem *sns*. It appears that the *gms* stem has been (incorrectly) inflected and mapped to the feature bundle of the *sns* stem. The correct inflection is *ysnendn*.

Generated Stem

The less frequent of the two are stems that have been randomly generated. For example with the stem given as *renzas*, the system generates: *ymryawem* in place of *yrenzawem*.

We have also encountered several looping errors such as: *ynawemaylmyylmyylmyylmyylmyylmyymayamawemyymamyamawemyymamyamawemyymamyamawemyymamyamawemyymamyamawemyylmyamyamawemyymamyamawemyymamyamawemyylmyylmyy* where the correct form is *ysnewem*.

5.2 Composition study

In Experiment 2, we tested the effects of training set composition; in other words, the informative nature of each verb type.

As mentioned above, the ambifixing verb class has the largest combinatorial space, reducing in size as we consider middle and prefixing verbs, respectively. Another way to consider this would be

[6] [M] marks the verb as middle and is present when one of the three middle prefixes is present.

	A&G	M&C	NNB
Ambifixing only	0.111	0.170	0.010
Middle only	0.121	0.210	0.111
Prefixing only	**0.212**	0.250	0.010
Ambi + Pre	0.111	0.190	0.010
Ambi + Mid	0.071	0.130	0.040
Mid + Pre	0.141	**0.290**	0.040
Ambi + Mid + Pre	0.061	0.200	0.040

Table 4: Data sets for each composition type, model and sampling accuracies. The training size for each is 386 forms (defined by the available prefixing verbs).

by providing comprehensive lists of the morphemes in a given language (such as Bickel and Nichols (2005); Shosted (2006)). Thus, the complexity of an inflectional system is measured by enumerating the number of inflectional categories and the range of available markers for their realisation (i.e. E-complexity). The bigger the number, the more complex the resulting system is.[7] With this in mind, we would expect that, given the same training size for each verb type, the ambifixing would perform the worst,[8] then the middle followed by the prefixing verbs. Our results, shown in Table 4, confirm this hypothesis.

More revealing than the overall accuracy for each set and model combination, is a decomposition of accuracy according to the verb class. Table 5 summarises the performance for each category according to verb class. Unsurprisingly, when the training set contains only one type of verb, it performs best for the type of verb seen in the training data.

From a linguist perspective, with principle parts from the middle verbs (mainly the suffixal system, recall that the middle verb takes a dummy prefix to reduce valency) and prefixing verbs (prefixal paradigm) we can construct the full paradigm available to ambifixing verbs. The results presented here show no such compositionality; instead, we see a simple correspondence to verb type observed.

As expected, we see the weak *leaking* or overlap between ambifixing and middle verbs, with very little transferability from prefixing to other verb types. It highlights the importance of tag choice; middle verbs have a [M] tag for the undergoer prefix, to mark the dummy prefix. If this tag were absent, would we see more transferability between ambifixing and middle verbs? Linguistically, no information would be lost as the absence of this tag still allows for the middle verbs to be clustered together.

5.3 Syncretism test

Experiment 3, entailed testing the systems with an unseen feature bundle and analysing the predicted forms, to gauge whether the models learnt syncretic behaviour.

As can be seen by the suffixal paradigm found in Evans (2016),[9] where both numbers are available, almost all the TAM categories exhibit syncretism across the second and third-person singular actor. The past perfective slot is the only case with distinct forms for the second and third singular person numbers. We are testing the prediction of an exception. The second singular is formed with {-nd-∅-} and the third person singular with the {-nd-a} suffix. We note the similarity between the second singular and dual forms, where the second dual is {-a-nd}. This becomes particularly pertinent when a vowel is inserted between consonants for ease of articulation but must also adhere to vowel harmony. In such cases, the second dual and second singular may appear the same.

Using the Aharoni and Goldberg (2017) architecture, the model incorrectly predicts 81 out of the 100 test forms as the third singular perfective category with the suffix {-nd-a} instead of {-nd–∅-}. Four forms predicted correctly (likely due to the similarity between the surface forms of the second person dual and singular tags) and the remaining fifteen distributed across second person dual and plural actor of the same TAM category, second/third singular for the imperfective non-prehodiernal TAM category, and several instances of nonce inflections such as {-ngt} or {-ngw}.

Similarly, the Makarov and Clematide (2018) system overwhelmingly predicts the unseen second singular form to be syncretic with the third singular (90 out of the 100 forms are predicted as such). Of the remaining ten instances three are correct, four are incorrectly modelled as the imperfective imperative (yet given the prefixing series is α, the future imperative prefix is absent) and one of

[7] Although more recent works have explored the issues with E-complexity (Ackerman and Malouf, 2013), we use it here as a guiding principle and acknowledge that further work is required to make a more nuanced statement.

[8] The combinatorial space for a transitive verb is $1,740$ cells (Muradoglu et al., 2020)

[9] Table 23.14 (pg 563) and Table 23.16 (pg 565)

	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 5: Absolute number of correct predictions for each setup.

each: second/third imperfective non-prehodiernal, second/third neutral preterite or second dual past perfective.

From these results, it is clear that such systems not only observe patterns that are directly stipulated through annotation but also others that may be inferred from the data. It is important to note this behaviour, particularly in cases such as the one presented here as the verb corpus only entails two instances of the second singular past perfective.

6 Conclusion

Diversity representation of languages in NLP is vital to test the generalisations of models. We present the first-ever neural network-based analysis of Nen, the first representation of the Yam language family and to the best of our knowledge, of a Papuan language. Nen provides an interesting case study as it exhibits non-monotonic morphological mapping: distributed exponence.

We compare state-of-the-art models for morphological reinflection across various training sizes and two sampling methods: random and Zipfian. The results show no significant difference between sampling methods, and minor differences may be attributed to training set composition differences. In the Zipfian case, the prefixing verb types are over-represented as they are more frequent in natural speech. We provide extensive analysis of types of errors generated by each system and show that the most common error type is allomorphy errors; a misapplication of morphophonological rules, or feature category mappings. We introduce a new subcategory of error type: free variation, which is a consequence of the natural speech origins of the corpus.

We further explore composition effects by generating training sets with incremental distributions for the three verb classes noted. As expected, we found that the models trained with one class had higher prediction accuracy for that class. Across homogeneous compositions, the prefixing verb class performed the best. This is likely due to a smaller E-complexity – or more simply – a smaller combination of feature tags for which the system must learn mappings. Finally, we explore the likelihood of learning syncretic behaviour and using this as a predictor for an unseen feature bundle – the second singular past perfective. Overwhelmingly, the system incorrectly predicts syncretism with over 80% for the A&G system and 90% for the M&C system. These results highlight that these systems can infer patterns from the data sets provided. Although in our case the prediction of syncretism mirrors that of a human learner, there may be underlying, unwanted properties learnt from the data given, which calls for careful preparation of data and observation of output.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, pages 429–464.

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proc. EMNLP*, Hong Kong.

Balthasar Bickel and Johanna Nichols. 2005. Inflectional synthesis of the verb. *The world atlas of language structures*, pages 94–97.

Matthew J. Carroll. 2016. *The Ngkolmpu Language*. Ph.D. thesis, The Australian National University.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

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.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Qianji Di, Ekaterina Vylomova, and Tim Baldwin. 2019. Modelling Tibetan verbal morphology. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 35–40, Sydney, Australia. Australasian Language Technology Association.

Jason Eisner. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Nicholas Evans. 2015. Valency in Nen. In Andrej Malchukov, Martin Haspelmath, Bernard Comrie, and Iren Hartmann, editor, *Valency classes: A comparative handbook*, pages 1069–1116. Berlin: Mouton de Gruyter.

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. In *Handbook of Quantifiers in Natural Language: Volume II*, pages 571–607. Springer.

Nicholas Evans. 2019a. Nen dictionary. *Dictionaria*, pages 1–5005.

Nicholas Evans. 2019b. *Waiting for the Word: Distributed Deponency and the Semantic Interpretation of Number in the Nen Verb*, pages 100–123. Edinburgh University Press.

Nicholas Evans. 2020. Waiting for the word: distributed deponency and the semantic interpretation of number in the Nen verb. In Andrew Hippisley Matthew Baerman, Oliver Bond, editor, *Morphological perspectives*, pages 100–123. Edinburgh: Edinburgh University Press.

Nicholas Evans and Julia Colleen Miller. 2016. Nen. *Journal of the International Phonetic Association*, 46(3):331–349.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

Alice C Harris. 2017. *Multiple exponence*. Oxford University Press.

Ronald M Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production*, volume 11. University of Helsinki, Department of General Linguistics Helsinki, Finland.

Vladimir I Levenstein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

John Mansfield. 2019. *Murrinhpatha morphology and phonology*, volume 653. Walter de Gruyter GmbH & Co KG.

Peter H Mathews. 1974. *Morphology: an introduction to the theory of word-structure*. Cambridge, England: Cambridge University Press.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311.

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.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting finite-state transductions with neural context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 623–633, San Diego, California. Association for Computational Linguistics.

Ryan K Shosted. 2006. Correlating complexity: A typological approach. *Linguistic Typology*, 10(1):1–40.

Katsuhito Sudoh, Shinsuke Mori, and Masaaki Nagata. 2013. Noise-aware character alignment for bootstrapping statistical machine transliteration from bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 204–209.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Chapter 4

Corpus and Model-Based Accounts of Paradigm Coverage

This chapter was published as:

Muradoglu, S., Suominen, H., & Evans, N. 2023. A Quest for Paradigm Coverage: The Story of Nen. In Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 74–85, Dubrovnik, Croatia. Association for Computational Linguistics.

Author contributions: S.M. and N.E. designed research; S.M. performed research; N.E. and S.M. contributed linguistic data; S.M., H.S., and N.E. analysed data; S.M performed data visualisation, and S.M. wrote the paper and revised it, based on critical comments by H.S. and N.E..

Chapter 3 revealed that neural network-based approaches can generalise paradigmatic patterns to unseen cells of the paradigm. It also hinted at the importance of training data composition. This raises the question of how well each part of the paradigm is represented and whether the under-represented morphosyntactic features can be predicted from those represented in the corpus. This chapter presents and contrasts a corpus-based account with a neural network-based account.

Language documentation aims to collate a collection representative of the language. The question of how to measure the completeness of the collection remains a topic of lively debate in the documentary linguistics literature. Himmelmann (1998) states that the purpose of language documentation is *'to provide a comprehensive record'*, yet as Bird (2015) questions, what constitutes a comprehensive record. In honour of this discussion, Baird et al. (2022) term this problem as the *'Himmelmann-Bird'* problem. This is the language documentation manifestation of a more general problem in statistics – sample representation[1] (Biber, 1993; Lüpke, 2009).

This paper explores this question using the bridge between corpus and computational linguistics. Given the enormity of the task of language documentation, the focus is narrowed down to Nen's verbal morphology. Model paradigm saturation[2] is measured by prompting the trained model to inflect a full paradigm for each verb type[3]. The accuracy obtained is used as a metric for paradigm coverage. To provide an empirical answer to the question of 'how much data is needed for X accuracy/coverage', a learning curve (accuracy as a function of annotation units) is modelled. The curve is modelled as a monotonic increasing function to present the best possible scenario. Specifically, the greater the number of annotation units considered, the greater the accuracy/paradigm coverage.

The corpus comprehensiveness is measured by following the trajectory of the most frequent verb for the type in question. The model is given a lemma at test time and asked to inflect the full paradigm to examine how well a model can capture generalisations. Four representative verbs are chosen for the test. The choice of these test verbs is motivated for several reasons: availability of verified paradigms, morphophonological similarity to the forms chosen for the corpus counterparts and regularity.

[1] Refer to 1.3.1 for further discussion.

[2] The term saturation in this sense is used interchangeably with coverage and comprehensiveness.

[3] Note that here type refers to the linguistic verb categories of transitive, middle, positional and copula, and not the type as defined in corpus or computational linguistics.

Recall that the experiments presented here involve the use of type-based datasets (discussed in more detail in 1.2.3). As such, the frequency of each wordform has no functional operation in the model or corpus-based account of coverage. Each wordform is encountered during train time once for the model-based account and is indexed by the corresponding lemma and MSDs (i.e., with no context). For the corpus-based account, the wordform is considered attested upon the first encounter, regardless of whether it is attested multiple times across the corpus. The available frequency measures from the specialised verb corpus by Muradoğlu (2017) are leveraged to sample training data in a similar manner described in chapter 3.

Moreover, the frequency of MSD combinations (actor/TAM and actor/undergoer) is utilised to establish the correlation between paradigm cell frequency (i.e., size of bubbles in figures) and model accuracy. We would expect that frequent MSD would be modelled with high accuracy. The results show that to be mostly true. The opposite does not appear to hold; less frequent MSDs are not necessarily modelled with low accuracy. For example, despite being represented in less than 1% of the corpus, the second person plural imperfective imperative ('IPFV.IMP:2plA, *{-ta-ng}*) or the first person singular neutral primitive ('NEUT.PRIM:1sgA', *{-tama-n}*)are modelled with high accuracy.

One explanation for this is the observed trade-off between paradigm size and irregularity in literature (Ackerman and Malouf, 2013; Cotterell et al., 2019b). Future, the well-known correlation between frequency and irregularity (Bybee, 1991), where irregular forms tend to be highly frequent in their usage, and low frequency forms tend to follow a regular pattern. As such a factor to consider is the complexity of the rule which governs the corresponding cell of the paradigm. Complexity here refers to allomorphy or conjugational differences. For example, the difference between the 'IPFV.BASIC:2|3sgA' form for *waprs* 'to make', *esrs* ' descend', *aebyängs* 'to fly' , *naprte, esrne* and *naednde* respectively. The results from this study do not necessarily correlate to inflection complexity. It can also reflect the representation of these various pattern

differences in the corpus.

It is no stretch that the needs of a model for generalisation can be distinct from that for archival purposes. For example, despite observing a second singular non-prehodiernal form for the verb *owabs* 'to speak', for archival purposes, the same morphosyntactic category is just as important for the verb *owans* 'set off'. From a modelling perspective, this is not entirely true. An unseen morphosyntactic category for the same verb *owabs* 'to speak' is potentially more informative (barring any phonotactic triggers for allomorphy). A side product of following the trajectory of coverage growth as a function of data size is the prediction afforded for data requirements to reach full coverage using a machine-learning-style learning curve. Following a common practice in ML applications, the data needed to achieve full coverage for the full verbal system and each verb type individually is extrapolated from the established learning curve.

The data size is reported with respect to units of a corpus, 'annotation units'. An annotation unit is defined as audibly demarcated units in the flow of speech (typically by pause breaks). Figure 4.1 shows three annotation units. For example, approximately five words must be encountered to encounter one verb. This average prediction is based on the statistics of the current Nen corpus, shown in figure 1.2. Although these units have been marked by hand in the corpus and are by no means strictly standardised, it characterises one of the common ways a linguist might annotate their recordings in ELAN. This unit is easily convertible to more standard metrics such as number of words or hours of recording. The motivation behind this unusual metric is to contextualise the data needs in a relevant way for field linguists.

This study examines four verb types: the copula, positional, middle and transitive verbs (See figure 1.4 for an overview of characteristics of each verb type). The results show that model coverage is significantly higher for the middle and transitive verbs compared to the coverage afforded by the corpus-based account. The positional verb is the one verb type, where the corpus coverage is higher than the model. This

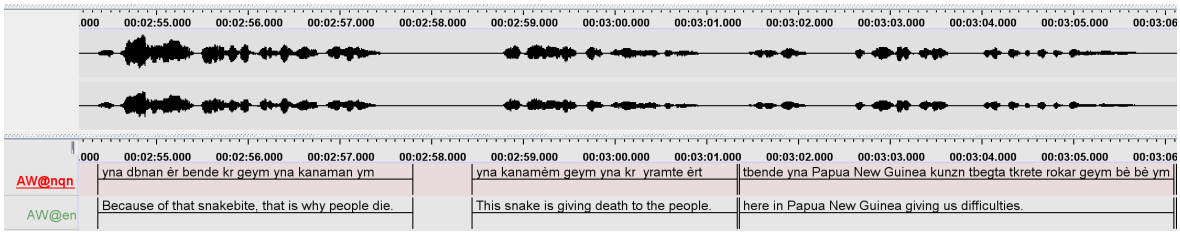


Figure 4.1: Example of three 'annotation units' or segments in ELAN for Nen. The top tier encodes the direct transcription of the audio file and the second tier shows a rough English translation.

finding is interesting, as the positional verb paradigm is the smallest of the three. It is likely due to the low representation of inflected positional verbs in the training set.

The purpose of an archive needs to be explicit, given that the distillation process for a task-oriented sub-corpus is highly costly. The results here suggest a transferability between data collected for language documentation. There is room to optimise the type of data if a particular output is desired, examples of this include active learning strategies[4].

4.0.1 MSD/gloss ablation study

Another question is the optimal number of tags to gloss the data. A small pilot study that did not make it into the paper is included. The focus is exploring whether the copula, positional, middle and transitive verbs should be marked as such (and thus treated as separate entities) or whether they should be minimally marked to allow leveraging from shared features. Given that, a field linguist and a language learner bootstrap the mapping of a linguistic paradigm by obtaining partial paradigms (either through elicitation or by natural means). Encountering a full paradigm for one verb is highly unlikely. Instead, the circumstantial context primes language informants to showcase verbs of different semantic domains. The field linguist most likely obtains part of the paradigm (either through elicitation or by natural means) for each verb. Depending on the degree of paradigmatic irregularity, these fragments may allow for a reconstruction of the full paradigm. With this in mind, we explore

[4]For general corpus building strategies see Barth and Schnell (2021).

		Verb Type				
Tags	**Metric**	Overall	Transitive	Middle	Positional	Copula
W/out [P],[C],[T],[M]	Acc	28.43	25.35	34.78	3.76	70.67
	Edit Distance	2.10	2.11	1.97	3.44	1.07
W/out [P],[C],[T]	Acc	27.72	24.62	42.55	16.54	71.33
	Edit Distance	2.08	2.23	1.92	3.35	1.09
With [P],[C],[T],[M]	Acc	30.81	26.31	49.07	0.75	68.00
	Edit Distance	2.45	2.62	2.00	2.98	1.09
W/out α, β, γ	Acc	4.69	0.06	4.04	1.50	60.67
	Edit Distance	3.75	4.06	3.30	3.71	1.27

Table 4.1: Transfomer model performance with varying degrees of gloss encoded in training data. [P], [C], [T], [M] note positional, copula, transitive and middle respectively. W/out is shorthand for 'without'. α, β and γ refer to series. See 1.1 for explanation.

whether allowing cross-transfer from one paradigm allows for greater coverage of the verbal morphological system.

The models are trained on the full corpus data (2,022 training samples) with the following variations in glossing:

- (1) Without any verb type tags
- (2) Without marking positional, copula and transitive explicitly. Middle verbs are marked. This option follows the Nen literature (as outlined in 1.1)
- (3) With all verb types marked
- (4) Without the α, β or γ series marked

Table 4.2 shows examples of the glossing for each experiment. Table 4.1 shows that the model does better overall when the verb types are not specified. However, the experiment setup in the paper follows the approach outlined by the Nen literature (examples shown in 1.1), experiment (2). On average, this approach yields the highest

	Verb Type			
	Copula	Positional	Middle	Transitive
Lemma	*m* 'to be'	*akingr* 'be standing (ndu)'	*owabs* 'to talk'	*yis* 'to plant'
Wordform	*ym*	*yakingr*	*nowabtan*	*yiwin*
(1)	V;IPFV.NPHD;3SGU;LGSPEC1	V;IPFV.NPHD;3SGU;LGSPEC1	V;IPFV.NPHD;1SGA;LGSPEC1	V;NEUT.PRET;1SGA;3SGU;LGSPEC1
(2)	V;IPFV.NPHD;3SGU;LGSPEC1	V;IPFV.NPHD;3SGU;LGSPEC1	V;IPFV.NPHD;1SGA;M;LGSPEC1	V;NEUT.PRET;1SGA;3SGU;LGSPEC1
(3)	V;IPFV.NPHD;3SGU;C;LGSPEC1	V;IPFV.NPHD;3SGU;P;LGSPEC1	V;IPFV.NPHD;1SGA;M;LGSPEC1	V;NEUT.PRET;1SGA;3SGU;T;LGSPEC1
(4)	V;IPFV.NPHD;3SGU	V;IPFV.NPHD;3SGU	V;IPFV.NPHD;1SGA;M	V;NEUT.PRET;1SGA;3SGU

Table 4.2: Example of MSD tags for each experiment across each verb type. Following unimorph (Kirov et al., 2018a) conventions for language specific phenomena, 'LGSPEC1,LGSPEC2,LGSPEC3' refer to α, β and γ respectively.

accuracies across the four verb types. The dataset, which tags verb type, benefits the transitive and middle verbs, but the performance of the prefixing verbs, particularly the positional verb, suffers. The absence of all verb type tags marginally benefits the transitive verb coverage but comes at a cost to the middle and positional verbs.

Lastly, given the ability of neural approaches to absorb implicit information, we examine whether the α,β,γ[5] dummy variables are needed or whether the model can infer co-requisite prefix suffix pairs to achieve particular TAM meanings. In some cases, the labelling of the series is not informative. For example, marking the γ for the imperfective remote past (see table 1.2) yields no additional information. At the same time, they are essential for other TAM categories, such as the imperfective non-prehodiernal (α) and imperfective yesterday past (β). Table 4.1 shows that the model performs poorly without these glosses.

It is important to note that the copula verb does consistently well across the glossing approaches as the model is exposed to a significant portion of the copula paradigm in training (given the large representation in the corpus).

Middle and transitive verbs share the suffixial system. Positional, transitive and copula verbs share the prefixial system. In the largest training set of 2,022 examples, 699 of the verb forms are middle, 51 are positional, 72 are copula, and 1200 are

[5]These are coded in data as 'LGSPEC1,LGSPEC2,LGSPEC3' in the data as per unimorph conventions.

transitive. All results reported for the main experiment follow the Nen literature glossing style (i.e., as in experiment (2). See table 4.2 for examples.). The main reason is the model performance.

These results suggest that the model is able to encode information beyond each example present in the corpus. In other words, the model can treat and infer information about a connected system (i.e., paradigm) rather than considering each inflection category as an independent string transduction.

Despite being a 'medium' resourced language (according to SIGMORHON terminology — i.e., the number of individual word-form samples available are in the 1,000s), the neural model performance is still quite low[6]. Additionally, it can be seen that the needs of corpus building can be different from that of an NLP system. The next section explores model intrinsic metrics (like model confidence scores) to guide sample collection. Prioritising NLP system performance is motivated by the benefits of automated glossing and processing. Further, the ability to capture generalisations allows for a compressed data quantity to capture the intricacies of the system at hand. Given the time pressures and resource costs of language documentation, this is an important point. To this end, the next chapter explores the idea of identifying informative examples in relation to model performance.

Often morphonological behaviour can be language-specific and not necessarily inflection feature specific. For example, if the model observes behaviour with one set of MSDs exhibiting vowel harmony, we expect the model to predict similar behaviour to an unrelated MSD. Certainly, the evidence from chapter 4 seems to support this hypothesis. So, is it enough to see a morphophonological pattern like vowel harmony in the frequent forms to extend the same alternative patterns to less frequent MSDs? This remains a direction for future research.

[6]It is possible to compare with other polysynthetic languages and the baselines reported in Goldman et al. (2023) (for example, Navajo obtains 52.1% accuracy for the same neural baseline). However, this should be taken as a very rough comparison, given the variation in linguistic complexity, data quality and the extent of the inflectional system included in the dataset.

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.

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

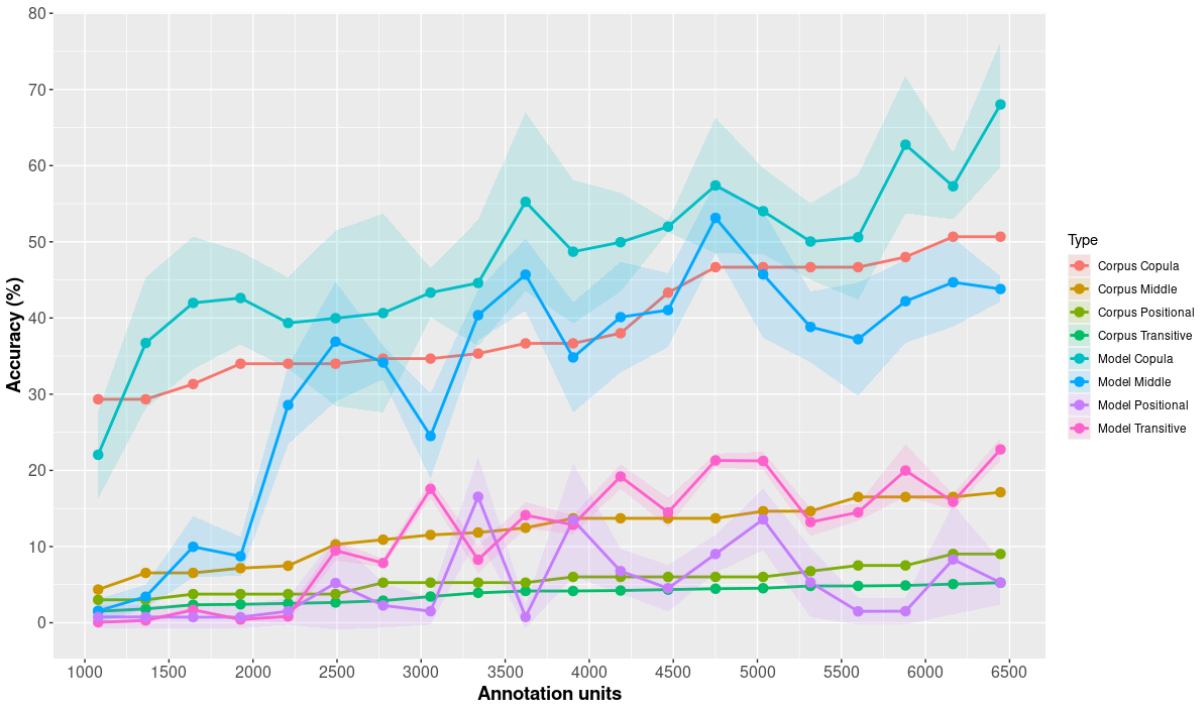


Figure 1: The coverage growth for four verb types in Nen, reported as a function of Annotation units (within corpus), where 'annotation units' are audibly-demarcated units in the flow of speech (typically by pause breaks). In our corpus, on average there is one verb per annotation unit, making annotation 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).

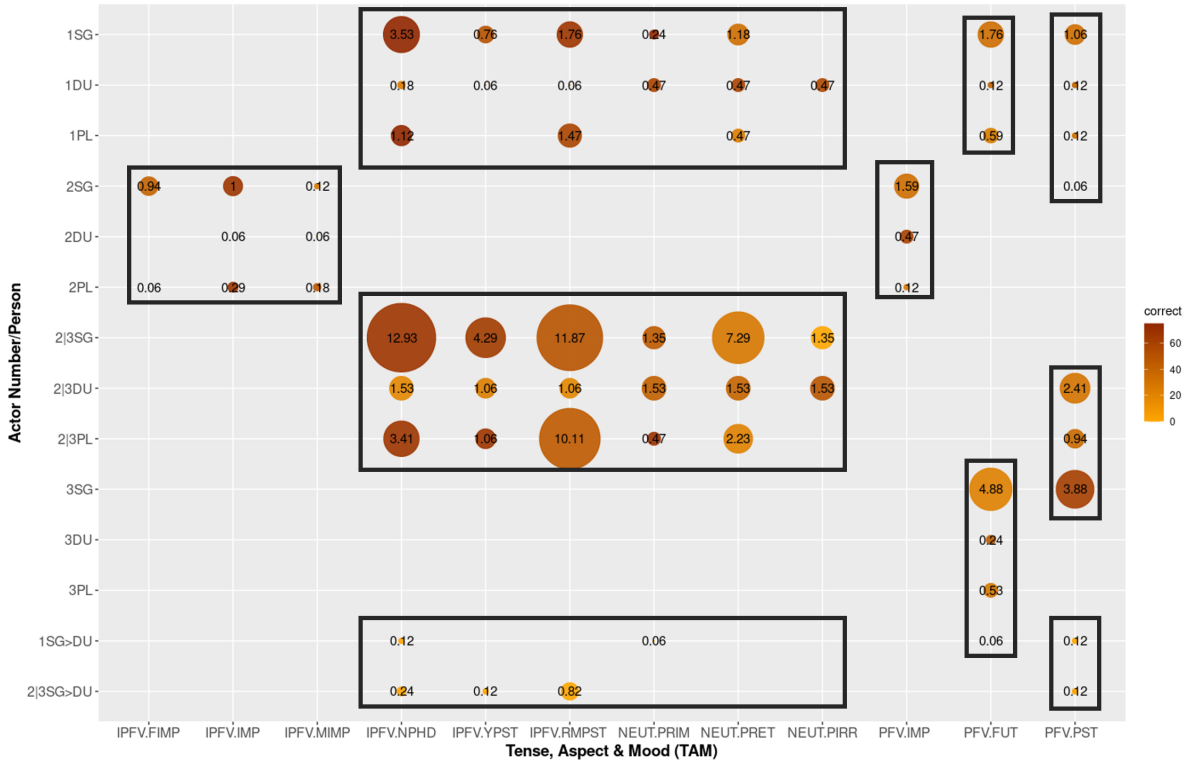


Figure 2: Bubble plot showcasing the frequency correlation between TAM and agent person number, reported numbers are percentage of corpus with the TAM/agent features. Navy lines indicate available cells described by the language design space. Note that the second and third persons are typically display 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

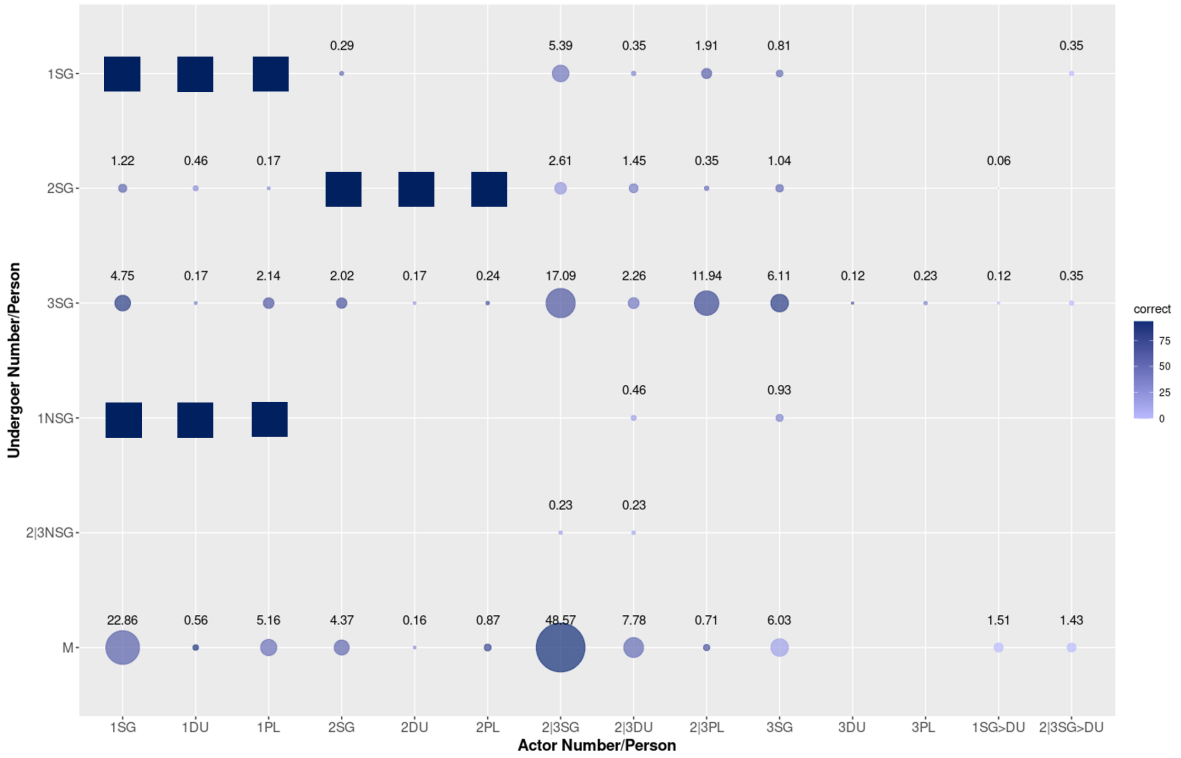


Figure 3: Relativised bubble plot of Actor and Undergoer person number for Nen. The navy blue blocks note the semantically disallowed combinations or in the case of first person acting on first person this meaning is achieved through reflexive constructions. The darker the colour (towards a purple) the 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

Chapter 5

Active Learning Guided Sample Collection

This chapter was published as:

Muradoglu, S., & Hulden, M. 2022. Eeny, meeny, miny, moe. How to choose data for morphological inflection.. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Author contributions: S.M. and M.H. designed research; S.M. performed research; S.M., M.H. analysed data; and S.M. wrote the paper and revised it based on critical comments by M.H..

Results from chapter 4 showed the interaction between high-frequency MSDs and model performance. As expected, in general, high-frequency MSDs were modelled with high accuracy. Notably, some low-frequency feature combinations were also predicted with high accuracy. These results highlight the discrepancy between the needs of the corpus for building an archival resource and model building. While an even sample distribution across each cell is regarded as ideal, this might not be the case for building NLP models. In particular, if the corresponding paradigm cell does

not exhibit a lot of allomorphy, the pattern the model must learn is simple. Where morphonological conjugation differences are present, the model must statistically infer the environment which conditions the allomorphs. Given these considerations, model-based methods to guide sample collection are examined. In this chapter, an active learning (AL) approach is explored to aid in prioritising linguistic annotation.

For example, consider the inflection patterns for the desinence suffix shown in 1.6. The pattern for the second singular person imperfective imperative ('2sgA:IPFV.IMP') is affixing *{-Ø}*. This rule is arguably more straightforward than the second singular person imperfective remote ('2sgA:IPFV.BASIC) expressed by displacing the previous vowel (*#e*). The latter pattern requires identifying the category of possible vowels. Another example is when vowel harmony effects are observed in some forms of the neutral preterite (Evans, 2016). Consider the transitive verb *wnḡis* 'to stand up', inflected for '2|3sgA:NEUT.PRET' is *ynḡiwi*, instead of *ynḡiwe*.

In this chapter, the focus is redirected from Nen and expanded to 30 typologically diverse languages. The primary reason for this is the quantity of data needed. Unfortunately, the current Nen corpus is smaller than the 3,500 training (5,000 in total) examples required for the experimental setup. An astute reader might question why a smaller training size was not considered. Despite advances in model performance for low-resource languages, the SIGMORPHON-defined low-resource (i.e., 100 training samples) setting is potentially prone to noise, whereby any data is beneficial. With this in mind, a relatively stable training size in terms of model performance has been chosen. Investigating AL with severely limited resources remains a question for future research.

Languages are chosen based on various WALS features pertaining to inflection, writing script, paradigm size and data availability. In addition to a one-cycle iteration for these languages, a detailed case study for the Austronesian language Natügu is presented. The reasons are severalfold; the data itself is IGT sourced, making it a more realistic scenario for language documentation, Natügu exhibits complex verbal

morphology, and it is a language of the Asia and Pacific.

As noted by Palmer (2009), active learning and label suggestion can reduce annotation costs, although the effectiveness is impacted by annotator expertise. The authors contrast three sampling methods: sequential, random, and uncertainty. Palmer (2009) which notes that sequential sampling methods perform worse by far, despite being the most common approach to annotation. Following this, sequential sampling is not considered in the experiments presented here. Instead, the remaining two methods (i.e., random and uncertainty-based sampling) is implemented. This chapter's reported model confidence-based sampling strategy is comparable to the Palmer (2009) uncertainty sampling.

The experiment presented in this chapter explores four sampling strategies for morphological inflection using a Transformer model. The first two sampling strategies are informed by model-intrinsic metrics: model confidence and entropy. The model confidence-based sampling relies on the transformer-generated loglikelihood associated with each form inflected. The higher the value, the more confident the model is of producing the correct form. From this, the correlation between correctness and model confidence is investigated. The results do not show a convincing correlation consistent across the 30 languages in this study. However, it should be noted that the four languages that derive data directly from IGTs are in the top eight Point-Biserial Correlation Coefficient (PBCC), showing a strong positive correlation between correctness and model confidence. The entropy-based sampling strategy extends the model confidence approach by treating the model calculated probabilities as a distribution rather than a single point.

Next, random selection is used to establish a comparable baseline. In other words, this baseline helps inform whether a guided sampling strategy is able to outperform random selection. For this method, the sampling process is repeated across three independent runs to account for variation and to report on a margin of uncertainty. As a measure of central tendency, the mean attempts to identify the central position

within the dataset. As such, if only two runs are considered, it is impossible to tell whether the returning two values are outliers of a similar type, polar opposites or representatives of the central point. The third run minimally allows for weighting to either of the first two runs. The margin of uncertainty also reports the variation of the actual results to the calculated mean. This accounts for the spread of accuracy observed across each run. While more runs would undoubtedly yield a more statistically stable value, the practical implications of computational costs (in both time and resources) meant limiting the independent runs to three.

The remaining strategy is informed by comparison to a gold standard. This strategy parallels the involvement of a linguist/language expert in the process, where the linguist/language expert can mark the model-generated form as either correct or incorrect. This is referred to as the oracle experiments in the paper. Symmetry is maintained by including conventionally disfavoured metrics. For example, high-confidence or correct forms are not typically resampled in AL experiments. Nevertheless, by including these metrics, data composition is shown empirically to be important. Despite conventional wisdom positing that larger data quantities yield better model performance, this is not the case when an incorrect abstraction is reinforced or other parts of the paradigm are queried at test time.

The underpinning theorem for this intuition is the law of large numbers. The law of large numbers is a theorem from probability and statistics, which states that the average of the results derived from repeating an experiment multiple times will better approximate the expected value. By extension, the more times the experiment is repeated, the closer the average is to the expected value. This idea extends to data in a machine-learning context. The training data should be representative and thus contain sufficient information to allow for generalisation to the unknown and underlying distribution of the population. From this, it is not hard to conceive that more data results in better performance. However, in low-resource settings, this approach is only sometimes feasible. Often, the corpora available are sparse for

morphologically complex languages (whether the inflectional system is synthetic, polysynthetic or agglutinating). As discussed earlier (section 1.1), a language such as Nen has multiple dimensions along which it encodes information. The multivariate nature of the distribution in question must be considered in these cases.

The results found a clear benefit for selecting data based on model confidence and entropy. Unsurprisingly, the oracle experiment, where only incorrectly inflected forms are chosen for further training, shows the most improvement. This is followed closely by choosing low-confidence and high-entropy predictions. As expected, random sampling sits between the two diametric ends of the metrics considered.

The experiment design is primarily motivated by possible integration into field linguistic methods and cycles. Before a field trip, a linguist typically prepares materials or questions for language consultants[1]. Each task or question typically focuses on one or two elements of a language. The envisaged usage would be training the model and generating predictions over a list of glosses. The list is then brought into the field to be evaluated by language informants[2]. The feedback is incorporated into the training data, and the model is retrained over a new set of examples. This process can be iterated until the model can reliably produce morphological glosses. For languages with large inflection tables (up to 1,740 wordforms for a transitive verb in Nen), it is unlikely to have many, if any, full inflectional tables for a lemma. It is also a taxing demand to task a linguist or native speaker with the production of such large tables. The ability to produce complete inflection tables is a valuable resource, not only for resource building in the form of aiding annotation but also for building pedagogical resources and language description – a common object found typically in the appendix of a grammar.

[1] For example, when first discovering the design space of a language, a linguist might start by creating a small list of vocabulary. The next step involves testing paradigmatic variables (such as subject person and number, object person and number, gender, TAM and polarity.) to get an initial feel for where there is inflectional complexity. After this, a linguist might design tasks around the inflectional complexity.

[2] A more detailed motivation behind implementing the morphological generation task is given in 1.2.3.

One possible future direction is to combine the approach in 4 with that presented here. That is utilising AL to address model coverage[3]. From a corpus-building perspective, it would make sense to prioritise sampling unseen inflected forms of lemma. However, results from 4 showed that the model can efficiently learn infrequently encountered inflection patterns. So, the model might be used to generate these forms.

Undoubtedly, a similar setup for morphological analysis would benefit the annotation bottleneck. A. McCarthy et al. (2019) outlines a task for in-context morphological analysis, where a sentence is given, and the task is to provide the lemma and MSD of each word. Another future research question might be to explore whether data for one direction (e.g., synthesis/generation) can be used in the counter direction (e.g., analysis). If a model achieves high performance in synthesis, can the same data (when the input and output pairs are inverted) be used to train for analysis? Would the performance be comparable? Can the model for morphological generation be used to generate data for morphological analysis?

[3]To an extent, the experiment presented here still concerns coverage, although not in the same systematic way presented in 4.

Eeny, meeny, miny, moe. How to choose data for morphological inflection

Saliha Muradoğlu[α,κ] **Mans Hulden**[χ]
[α]The Australian National University (ANU) [χ]University of Colorado
[κ]ARC Centre of Excellence for the Dynamics of Language (CoEDL)
saliha.muradoglu@anu.edu.au, mans.hulden@colorado.edu

Abstract

Data scarcity is a widespread problem in numerous natural language processing (NLP) tasks for low-resource languages. Within morphology, the labour-intensive work of tagging/glossing data is a serious bottleneck for both NLP and language documentation. Active learning (AL) aims to reduce the cost of data annotation by selecting data that is most informative for improving the model. In this paper, we explore four sampling strategies for the task of morphological inflection using a Transformer model: a pair of oracle experiments where data is chosen based on whether the model already can or cannot inflect the test forms correctly, as well as strategies based on high/low model confidence, entropy, as well as random selection. We investigate the robustness of each strategy across 30 typologically diverse languages. We also perform a more in-depth case study of Natügu. Our results show a clear benefit to selecting data based on model confidence and entropy. Unsurprisingly, the oracle experiment, where only incorrectly handled forms are chosen for further training, which is presented as a proxy for linguist/language consultant feedback, shows the most improvement. This is followed closely by choosing low-confidence and high-entropy predictions. We also show that despite the conventional wisdom of larger data sets yielding better accuracy, introducing more instances of high-confidence or low-entropy forms, or forms that the model can already inflect correctly, can reduce model performance.

1 Introduction

The need for linguistically annotated data sets is a drive that unites many fields within linguistics. Computational linguists often use labelled data sets for developing NLP systems. Theoretical linguists may utilise corpora for constructing statistical argumentation to support hypotheses about language or phenomena. Documentary linguists create interlinear glossed texts (IGTs) to preserve linguistic and cultural examples, which typically aids in generating a grammatical description. With the renewed focus on low-resource languages and diversity in NLP and the urgency propelled by language extinction, there is widespread interest in addressing this bottleneck.

One method for reducing annotation costs is active learning (AL). AL is an iterative process to optimise model performance by choosing the most critical examples to label. It has been successfully employed for various applications through NLP tasks including deep pre-trained models (BERT) (Ein-Dor et al., 2020), semantic role labelling (Myers and Palmer, 2021), named entity recognition (Shen et al., 2017), word sense disambiguation (Zhu and Hovy, 2007), sentiment classification (Dong et al., 2018) and machine translation (Zeng et al., 2019; Zhang et al., 2018). The iterative nature of AL aligns nicely with the language documentation process. It can be tied into the workflow of a field linguist who consults with a language informant or visits a field site in a periodic manner. Prior to a field trip, a linguist typically prepares material/questions (such as elicitation's or picture tasks[1]) for language consultants which may focus on elements of the language they are working to describe or for material creation (e.g., pedagogical). We propose AL as a method which can provide a supplementary line of insight into the data collection process, particularly for communities that wish to develop and engage with language technology and/or resource building.

Previous work by Palmer (2009) details the efficiency gains from AL in the context of language documentation for the task of morpheme labelling. With deep learning models leading performance for the task of morphological analysis (Pimentel et al., 2021; Vylomova et al., 2020; McCarthy et al.,

[1]Or indeed any materials such as those complied by the Max Planck Institute for Psycholinguistics at http://fieldmanuals.mpi.nl/

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7294 - 7303
December 7-11, 2022 ©2022 Association for Computational Linguistics

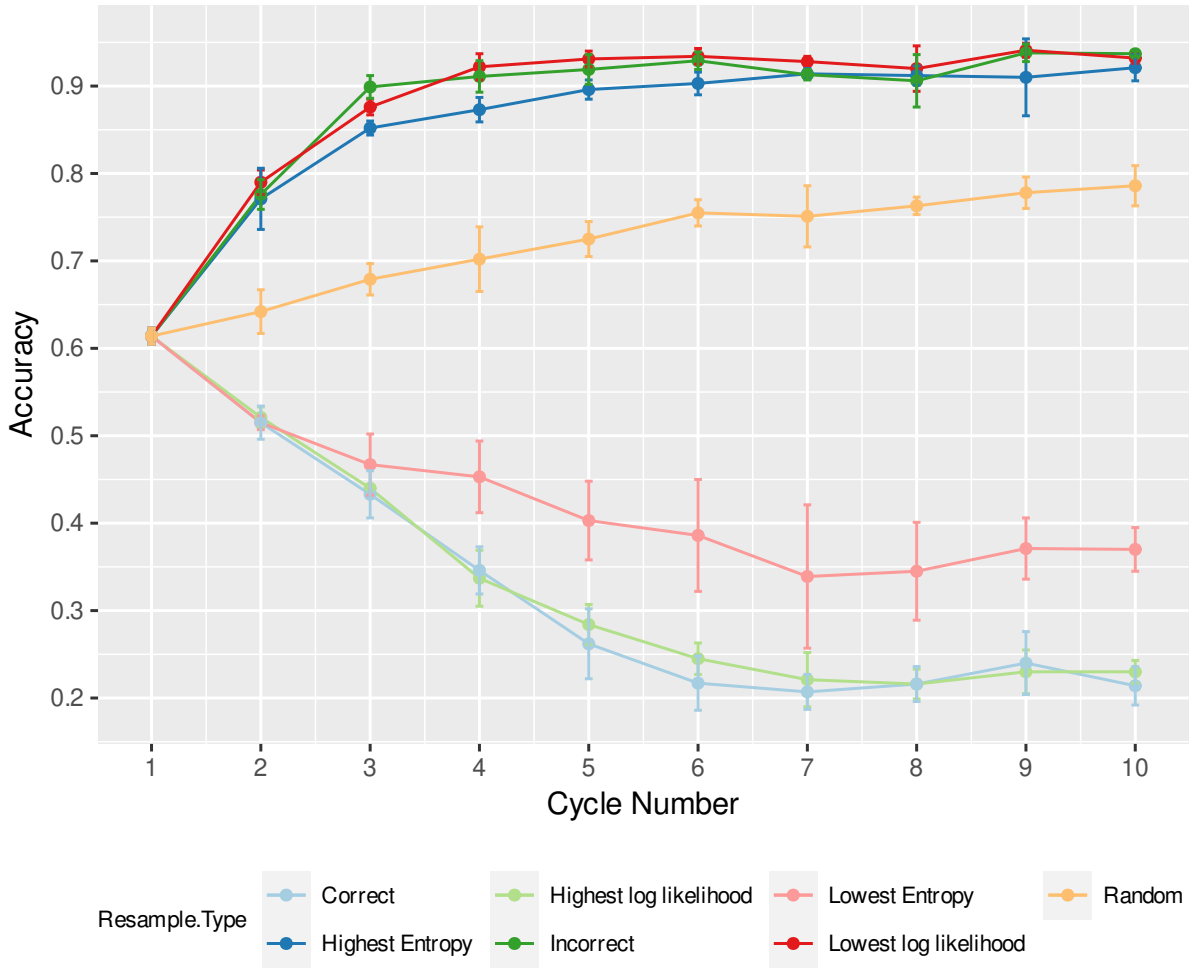


Figure 1: The accuracy for each trained modelled, starting from the baseline (cycle 1). Each cycle 250 instances are re-sampled via the seven sampling methods: correct/incorrect, high/low model confidence, high/low entropy and random (coded with colour). The reported error bars are calculated across 3 separate runs. See Table 1 in Appendix for more detail. After cycle 2, the same sampling strategy is applied to that stream of experiment - e.g. for the lowest log-likelihood strategy, from cycle 2 to 10 the same strategy is used.

2019), AL in the context of neural methods is needed.

This paper addresses the following question: How can we identify the type of data needed to improve model performance? To answer this, we explore the use of AL for the task of morphological inflection using a Transformer model. We run AL simulation experiments with four different sampling strategies: (1) correctness oracle, (2) model confidence, (3) entropy and (4) random selection. These strategies are tested across 30 typologically diverse languages and a 10-cycle iterative experiment using Natügu as a case study.

2 Data

We use data from the UniMorph Project (McCarthy et al., 2020), Interlinear Glossed Texts (IGT) from Moeller et al. (2020) and SIGMORPHON (Vylomova et al., 2020; Pimentel et al., 2021). In addition to the data availability, we consider typological diversity when selecting languages to include. Broadly, we attempt to include types of languages that exhibit varying degrees of complexity for inflection. We also consider morphological characteristics coded in WALS; prefixing vs. suffixing (Dryer, 2013), inflectional synthesis of the verb (Bickel and Nichols, 2013b) and exponence (Bickel and Nichols, 2013a). An additional consideration is the paradigm size for the morphological system modelled.

We note data source type to account for the variation in standard across Wikipedia, IGT field data, glossed examples from grammars and data generated from computational grammars.

3 Experimental Setup

We train the model as if we were addressing an 'inflection' task (Vylomova et al., 2020). The data is in the form of triplets: lexeme, morphosyntactic tags and the desired output inflected form (e.g. ⟨`walk, V;PST, walked`⟩)[2]. Each model is trained with the fairseq Transformer (Ott et al., 2019) and our hyperparameters follow Liu and Hulden (2020).

A baseline model is trained, after which more examples are resampled from the baseline test file using the methods detailed below. The initial baseline model is trained with 3,500 instances, 1,000 test and 500 for development. We resample 250 instances.

3.1 Sampling strategies

Oracle The oracle experiments serve as a proxy for linguist/language expert feedback. 250 examples are sampled based on whether the predicted form is correct/incorrect. The initial filter is supplemented with the following criteria: (1) if there are fewer than 250 incorrect forms, the remaining slots are filled in accordance with examples that exhibit the smallest difference between the first and second output form's log-likelihood, (2) in the case of more than 250 incorrect forms, the incorrect instances are ranked based on the maximum Levenshtein distance between the predicted and target forms. The same selection criteria are applicable for the counterpart correct experiment, with reversed limits (e.g. in the case of less than 250 correct forms, the instances with the largest difference between the first and second log-likelihood are considered).

Model Confidence The instances introduced to the training data are sampled based on the model confidence for each form. In this particular strategy, we only record the log-likelihood for the highest-ranked prediction in the beam.

We further examine the correlation between the log-likelihood (continuous variable) and accuracy (dichotomous variable) of the best prediction generated by the model by calculating the Point-Biserial Correlation Coefficient (PBCC). Across the 30 languages we study, the average PBCC is 0.388. Like all correlation coefficients, the PBCC measures the strength of the correlation, and the reported value ranges from -1 to +1, where -1 indicates an inverse association, +1 indicates a positive association, and 0 indicates no association at all.

[2]Data and code available at https://github.com/smuradoglu/ALmorphinfl

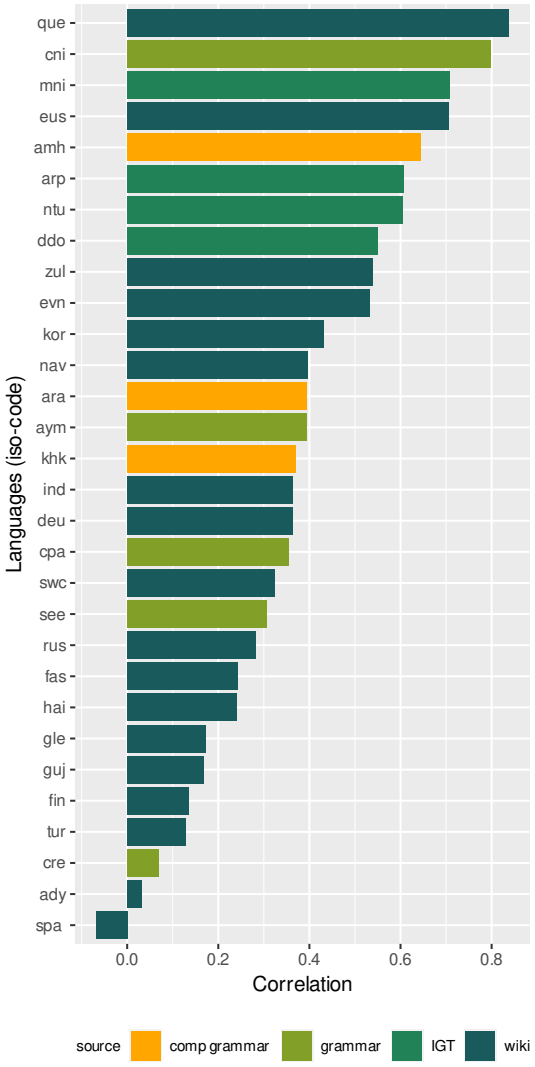


Figure 2: The calculated Point-Biserial Correlation Coefficient (PBCC) between correct prediction and the model log-likelihood, across 30 different languages. The source of the data is also noted with colour.

Entropy Here we expand upon the previous strategy—model confidence. We consider the distribution of the ranked output predictions for a particular input and approximate its entropy $-\sum_i p_i \log(p_i)$, by only considering such predictions where $p_i \geq 0.05$, i.e. we calculate $-\sum p_i \log(p_i)$, for all $p_i \geq 0.05$. The model generated log-likelihoods are converted to probabilities and renormalised across the outputs generated by beam search. $p_i = \frac{p_n}{\sum_{j=1}^{b} p_n}$, b being the number of predictions we retrieve from the beam search.

Random We contrast the previous methods for re-sampling with random data selection. To establish whether the change in accuracy is statistically significant, we report the average across three independent runs and the standard deviation across the measured accuracy.

4 Results and Discussion

To simulate a documentation process, we have chosen Natügu as a case study. The inflection data is from Moeller et al. (2020) and is derived from IGTs—a form that is commonly utilised by field linguists. Our choice of language is further motivated by the morphological complexity exhibited by Natügu. By all accounts Natügu showcases complex morphology (Wurm, 1976; Åshild Næss and Boerger, 2008), particularly on the verb. Historically, this observed complexity led to the language family named as Papuan instead of Austronesian.

Additionally, we observe a positive correlation between prediction correctness and model confidence (0.605). In fact, 4 out of the top 8 correlations (as shown in Figure 2) are languages with IGTs as a data source. For these reasons, we have chosen to examine iterative sampling over 10 cycles.

Figure 1 summarises our results for Natügu. The re-sampling process is iterated over 10 cycles. The first cycle is the baseline/seed run and consists of a 600 instance training set. To account for the impact of random factors affecting the initial training data selection, we have conducted 3 independent seed runs—differing solely on the initial training set. The average accuracy and corresponding standard deviation is reported with the error bars.[3]

The small starting size is motivated by the parallels with language documentation efforts, which are typically a low-resource setting. In each cycle, 250 forms are sampled via the corresponding sampling strategy. By the last cycle the training data consists of 2,850 instances.

Aside from the 3rd and 10th cycle, the lowest log-likelihood sampling consistently provides the greatest improvement. For these two cycles sampling based on incorrect forms outperforms selection based on low confidence. In general, the top 3 selection methods are ranked as follows: low log-likelihood, incorrect and highest entropy forms. We note the possible interplay between paradigm size (907 unique tag combinations) and training size set (1,100 by cycle 3); unseen morphosyntactic categories will be most informative and presumably beneficial to model performance.

Given the strong correlation between prediction accuracy and model confidence for Natügu, we expect similarity in trajectory across cycle number and accuracy for the oracle and model-confidence based sampling strategies. Figure 1 verifies these forecasts; we see that the sampling based on prediction correctness (in light blue) and the sampling based on the highest log-likelihood (in light green) almost look identical. The same is observable for low log-likelihood (in red) and sampling based on incorrect prediction (green).

The lowest log-likelihood sampling method can be seen as an approximation for the highest entropy selection method, and by extension, the highest log-likelihood as an approximation for the lowest entropy selection. Our results for iterative AL for Natügu show that choosing by approximation is a higher risk endevaour. The choice either works really well or not at all. When we contrast low entropy and high model confidence as a selection strategy we can see that low entropy limits the impact of high model confidence since it accounts for a distribution rather than the single value approximation. We observe similar behaviour between the the high entropy and low confidence selection strategies. Random sampling shows gradual improvement.

Work by Yuan et al. (2020) highlight the issues with uncertainty sampling for deep learning models; noting that neural networks are poorly calibrated (Guo et al., 2017), and that the correlation between high confidence and correctness is not well established. We explore this correlation for our models in Figure 2. We observe a similar uncertainty with an overall slight positive correlation across the 30 languages examined. Despite this, our results show that data selection based on low model confidence yields significant improvement of model accuracy. The work presented here is intended as a preliminary baseline; we leave it to future work to consider calibration methods such as temperature scaling.

Interestingly, despite an increase in training data size, introducing new data that the model already can inflect correctly, or low-entropy or high-confidence forms actually reduces model performance despite the widely-held notion that more data is better. Another recent study by Samir and

[3] Individual values can be found in Table 1 of the Appendix.

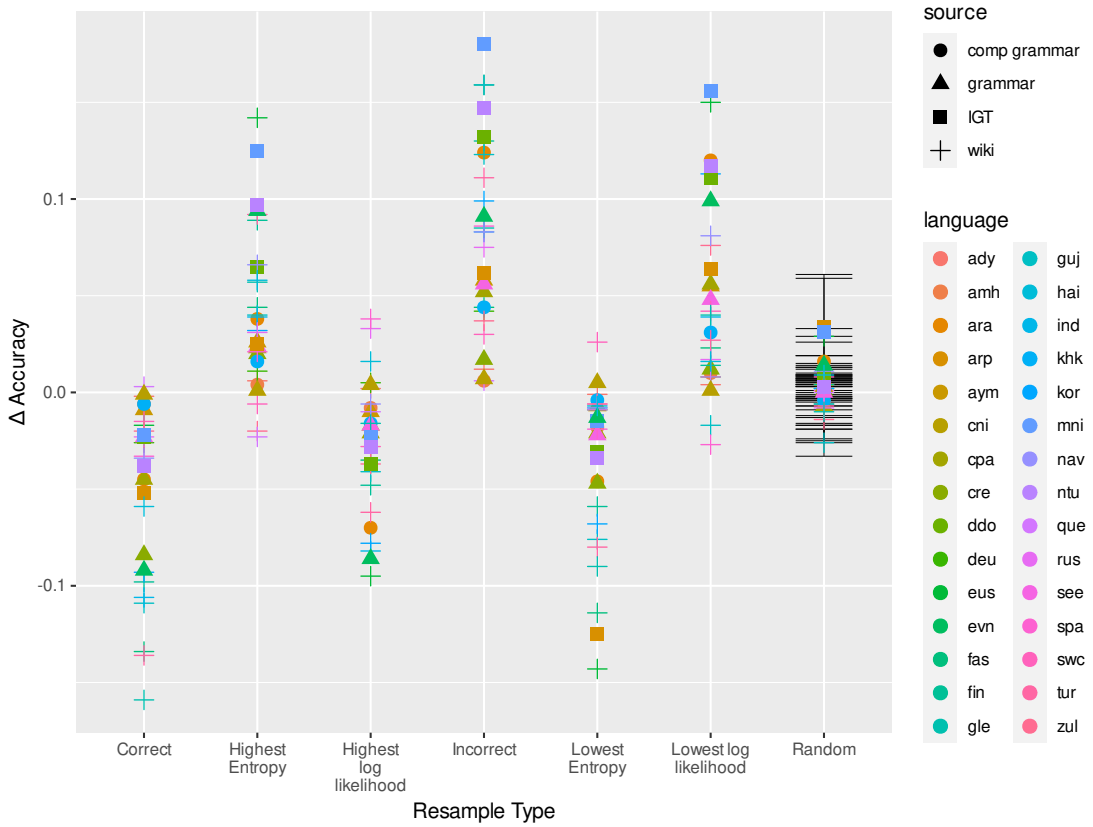


Figure 3: The change in accuracy (from the established baseline) is reported with each sampling strategy, across 30 different languages (coded with colour). The source of data is also noted with tick shapes.

Silfverberg (2022) reports similar behaviour, where data hallucination reduces prediction accuracy for words that exhibited reduplication.

We extend the same sampling strategies to 30 different languages for one round of re-training. The results are summarised in Figure 3. Within the 30 languages we ensure to include languages with large inflection table sizes (ranging from 12 to 700+), different scripts (Latin, Cyrillic, Arabic, Hangul, Ge'ez and Gujarati) and morphological typology (agglutinating, fusional, polysynthetic). We code for the source of the data, and see no particular deviation from the overall observed behaviour. The reported error bars for random sampling correspond to the standard deviation across three independent runs of random sampling.

It is clear that in general, the sampling strategies can be ordered for prediction accuracy improvement in the following manner: incorrect, lowest log-likelihood, highest entropy, random, highest log-likelihood, lowest entropy and finally correct form sampling. While a handful of languages deviate from this pattern (e.g. Swahili or Dido),[4] it holds true for a majority of the languages considered.

5 Conclusion

In this paper we examine four different sampling strategies within an AL framework for modelling morphological inflection using a Transformer model. We consider correct/incorrect prediction, model confidence, entropy and random selection as sampling strategies. Our results clearly show that AL can significantly improve learning rates for morphological inflection. Unsurprisingly, adding oracle-indicated incorrect forms for training yields the greatest model improvement. In the absence of a language expert, model confidence can be used to prioritise data annotation. This holds true across 30 different languages. We also show that larger datasets do not always yield better results; the diversity of the training set matters.

Future research should extend the analysis to incorporate language-specific factors—such as model performance for each morphosyntactic slot within the morphological paradigm.

[4] see Table.3 in Appendix for more detail.

Limitations

The primary limitation of this study is that the results are not evaluated in a real life documentation scenario. While we have tried to address this gap by noting the source of data, and have enlisted IGT data to serve as a proxy, we acknowledge that fieldwork data is often inconsistent, noisy and requires much more data cleaning. The data used for these experiments is, for the most part, already structured as a paradigm.

In addition, the simple metric of accuracy can be crude and is often prone to some degree of fluctuation. To minimise these effects we have considered the change in accuracy across sampling cycles instead. Lastly, we have tried to collate a diverse set of languages to consider. However, this is largely limited by the availability of data. It is likely that several morphophonological phenomena are not included within the data sets used here.

References

Balthasar Bickel and Johanna Nichols. 2013a. Exponence of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Balthasar Bickel and Johanna Nichols. 2013b. Inflectional synthesis of the verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.

Matthew S. Dryer. 2013. Prefixing vs. suffixing in inflectional morphology. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Skatje Myers and Martha Palmer. 2021. Tuning deep active learning for semantic role labeling. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 212–221, Groningen, The Netherlands (online). Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Mary Palmer. 2009. *Semi-automated annotation and active learning for language documentation*. Ph.D. thesis.

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.

Farhan Samir and Miikka Silfverberg. 2022. One wug, two wug+s: Transformer inflection models hallucinate affixes. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 31–40, Dublin, Ireland. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Stephen A. Wurm. 1976. The Reef Islands-Santa Cruz family. In Stephen A. Wurm, editor, *New Guinea Area Languages and Language Study Vol 2: Austronesian Languages*, volume 39 of *Pacific Linguistics: Series C*, pages 637–674. Research School of Pacific and Asian Studies, Australian National University, Canberra.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Empirical evaluation of active learning techniques for neural MT. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, Hong Kong, China. Association for Computational Linguistics.

Pei Zhang, Xueying Xu, and Deyi Xiong. 2018. Active learning for neural machine translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 153–158.

Jingbo Zhu and Eduard Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790, Prague, Czech Republic. Association for Computational Linguistics.

Åshild Næss and Brenda H. Boerger. 2008. Reefs-Santa Cruz as Oceanic: Evidence from the verb complex. *Oceanic Linguistics*, 47(1):185–212.

A Appendix

		Resample by:														
		Lowest $log(p_i)$					Highest $log(p_i)$					Random				
Cycle #	training size	S1	S2	S3	avg	std	S1	S2	S3	avg	std	S1	S2	S3	avg	std
1	600	0.618	0.621	0.604	0.614	0.009	0.618	0.621	0.604	0.614	0.009	0.618	0.621	0.604	0.614	0.009
2	850	0.800	0.795	0.774	0.790	0.014	0.508	0.532	0.522	0.521	0.012	0.617	0.666	0.642	0.642	0.025
3	1100	0.870	0.872	0.886	0.876	0.009	0.439	0.442	0.439	0.440	0.002	0.679	0.697	0.662	0.679	0.018
4	1350	0.927	0.933	0.905	0.922	0.015	0.374	0.317	0.321	0.337	0.032	0.723	0.723	0.659	0.702	0.037
5	1600	0.922	0.939	0.932	0.931	0.009	0.275	0.310	0.267	0.284	0.023	0.727	0.744	0.705	0.725	0.020
6	1850	0.934	0.925	0.943	0.934	0.009	0.266	0.234	0.236	0.245	0.018	0.741	0.771	0.754	0.755	0.015
7	2100	0.921	0.929	0.933	0.928	0.006	0.256	0.198	0.208	0.221	0.031	0.735	0.726	0.791	0.751	0.035
8	2350	0.940	0.890	0.929	0.920	0.026	0.236	0.207	0.206	0.216	0.017	0.756	0.774	0.758	0.763	0.010
9	2600	0.943	0.932	0.948	0.941	0.008	0.225	0.208	0.258	0.230	0.025	0.758	0.783	0.794	0.778	0.018
10	2850	0.929	0.927	0.939	0.932	0.006	0.242	0.230	0.217	0.230	0.013	0.760	0.795	0.803	0.786	0.023

Table.1: Model accuracies for iterative sampling for Natügu, across the lowest and highest low-likelihoods and random sampling strategies. S1, S2, S3 corresponds to seed 1, seed 2 and seed 3 respectively. Avg and std indicate the average value across the three seed runs and the standard deviation. Data used to generate Figure.1.

		Resample by:																			
		Incorrect					Correct					Highest Entropy					Lowest Entropy				
Cycle #	training size	S1	S2	S3	avg	std	S1	S2	S3	avg	std	S1	S2	S3	avg	std	S1	S2	S3	avg	std
1	600	0.618	0.621	0.604	0.614	0.009	0.618	0.621	0.604	0.614	0.009	0.618	0.621	0.604	0.614	0.009	0.618	0.621	0.604	0.614	0.009
2	850	0.778	0.791	0.758	0.776	0.017	0.530	0.521	0.493	0.515	0.019	0.792	0.790	0.731	0.771	0.035	0.506	0.520	0.518	0.515	0.008
3	1100	0.896	0.887	0.913	0.899	0.013	0.452	0.446	0.402	0.433	0.027	0.848	0.861	0.848	0.852	0.008	0.507	0.445	0.448	0.467	0.035
4	1350	0.901	0.931	0.900	0.911	0.018	0.373	0.346	0.320	0.346	0.027	0.876	0.857	0.885	0.873	0.014	0.500	0.439	0.421	0.453	0.041
5	1600	0.923	0.934	0.901	0.919	0.017	0.291	0.278	0.216	0.262	0.040	0.892	0.908	0.888	0.896	0.011	0.454	0.386	0.369	0.403	0.045
6	1850	0.935	0.935	0.917	0.929	0.010	0.212	0.189	0.250	0.217	0.031	0.895	0.895	0.918	0.903	0.013	0.460	0.342	0.357	0.386	0.064
7	2100	0.906	0.916	0.916	0.913	0.006	0.189	0.229	0.203	0.207	0.020	0.919	0.916	0.908	0.914	0.006	0.362	0.248	0.408	0.339	0.082
8	2350	0.929	0.872	0.917	0.906	0.030	0.234	0.220	0.194	0.216	0.020	0.919	0.917	0.899	0.912	0.011	0.408	0.301	0.327	0.345	0.056
9	2600	0.928	0.947	0.940	0.938	0.010	0.210	0.229	0.280	0.240	0.036	0.933	0.860	0.938	0.910	0.044	0.365	0.409	0.339	0.371	0.035
10	2850	0.935	0.940	0.937	0.937	0.003	0.229	0.188	0.224	0.214	0.022	0.933	0.904	0.925	0.921	0.015	0.378	0.389	0.342	0.370	0.025

Table.2: Model accuracies for iterative sampling for Natügu, across incorrect, correct, highest and lowest entropy sampling strategies. S1, S2, S3 corresponds to seed 1, seed 2 and seed 3 respectively. Avg and std indicate the average value across the three seed runs and the standard deviation. Data used to generate Figure.1.

Language	Iso-code	PBCC	p-value
Adyghe	ady	0.031	3.29E-01
Amharic	amh	0.643	5.74E-118
Arabic	ara	0.394	1.53E-38
Arapaho	arp	0.607	1.08E-101
Aymara	aym	0.394	1.64E-38
Asháninka	cni	0.799	1.10E-222
Palantla Chinantec	cpa	0.355	4.10E-31
Cree	cre	0.069	2.91E-02
Dido	ddo	0.550	4.24E-80
German	deu	0.363	1.60E-32
Basque	eus	0.707	3.67E-152
Evenki	evn	0.532	2.94E-74
Persian	fas	0.242	7.68E-15
Finnish	fin	0.135	1.84E-05
Irish	gle	0.172	4.24E-08
Gujarati	guj	0.168	8.40E-08
Haida	hai	0.240	1.47E-14
Indonesian	ind	0.364	1.05E-32
Halh Mongolian	khk	0.370	8.44E-34
Korean	kor	0.431	1.78E-46
Manipuri	mni	0.709	2.34E-153
Navaho	nav	0.396	7.66E-39
Natügu	ntu	0.605	1.13E-100
Quechua	que	0.838	3.44E-265
Russian	rus	0.282	9.57E-20
Seneca	see	0.306	3.75E-23
Spanish	spa	-0.069	3.02E-02
Swahili	swc	0.324	6.16E-26
Turkish	tur	0.129	4.17E-05
Zulu	zul	0.540	9.23E-77

Table.2: Correct and model log-likelihood correlation based on baseline for each language. The reported value is a Point-Biserial Correlation Coefficient (PBCC) with the respective p-value. Data used to generate Figure.2.

				Resample by:										
Language	Source	# Tables	Baseline	Correct	Incorrect	Lowest $log(p_i)$	Highest $log(p_i)$	Lowest Entropy	Highest Entropy	$Random_1$	$Random_2$	$Random_3$	$Random_{avg}$	± Std dev
ady	Wiki	430	0.986	0.984	0.998	0.990	0.988	0.985	0.992	0.989	0.990	0.990	0.990	0.001
amh	c. grammar	285	0.983	0.977	0.989	0.993	0.975	0.965	0.987	0.980	0.977	0.977	0.978	0.002
ara	c. grammar	83	0.800	0.755	0.924	0.920	0.730	0.754	0.838	0.818	0.805	0.824	0.816	0.010
aym	grammar	55	0.933	0.924	0.991	0.988	0.923	0.926	0.959	0.939	0.941	0.938	0.939	0.002
cpa	grammar	490	0.843	0.798	0.895	0.899	0.822	0.822	0.866	0.820	0.857	0.832	0.836	0.019
cre	grammar	22	0.113	0.029	0.130	0.125	0.096	0.066	0.133	0.110	0.116	0.115	0.114	0.003
deu	wiki	450	0.937	0.911	0.979	0.945	0.942	0.928	0.948	0.935	0.939	0.920	0.931	0.010
eus	wiki	12	0.755	0.738	0.914	0.905	0.660	0.612	0.897	0.813	0.754	0.784	0.784	0.030
fas	wiki	39	0.178	0.044	0.222	0.201	0.143	0.064	0.222	0.186	0.182	0.178	0.182	0.004
fin	wiki	97	0.587	0.489	0.717	0.601	0.539	0.528	0.676	0.593	0.584	0.590	0.589	0.005
guj	wiki	280	0.620	0.511	0.743	0.603	0.579	0.544	0.660	0.601	0.587	0.594	0.594	0.007
ind	wiki	750	0.551	0.445	0.634	0.590	0.469	0.543	0.590	0.556	0.530	0.549	0.545	0.013
khk	c. grammar	720	0.936	0.930	0.980	0.967	0.920	0.932	0.952	0.944	0.934	0.918	0.932	0.013
kor	wiki	60	0.597	0.504	0.696	0.710	0.519	0.529	0.629	0.595	0.597	0.606	0.599	0.006
rus	wiki	320	0.884	0.861	0.959	0.901	0.917	0.878	0.915	0.857	0.889	0.881	0.876	0.017
see	grammar	135	0.895	0.872	0.951	0.943	0.878	0.873	0.919	0.902	0.884	0.898	0.895	0.009
spa	wiki	75	0.880	0.847	0.966	0.853	0.918	0.861	0.901	0.884	0.874	0.884	0.881	0.006
swc	wiki	53	0.931	0.916	0.961	0.973	0.903	0.957	0.925	0.939	0.941	0.937	0.939	0.002
tur	wiki	35	0.464	0.328	0.575	0.491	0.402	0.384	0.556	0.456	0.462	0.452	0.457	0.005
zul	wiki	62	0.881	0.861	0.918	0.957	0.844	0.875	0.861	0.876	0.868	0.856	0.867	0.010
arp	IGT	470	0.290	0.238	0.352	0.354	0.265	0.165	0.315	0.326	0.296	0.349	0.324	0.027
que	wiki	25	0.982	0.985	0.988	0.990	0.972	0.973	0.959	0.969	0.994	0.982	0.982	0.013
gle	wiki	350	0.387	0.228	0.472	0.427	0.371	0.297	0.444	0.372	0.375	0.385	0.377	0.007
ddo	IGT	400	0.793	0.770	0.925	0.904	0.756	0.762	0.858	0.804	0.799	0.806	0.803	0.004
nav	wiki	280	0.860	0.826	0.943	0.941	0.854	0.852	0.926	0.874	0.862	0.877	0.871	0.008
mni	IGT	525	0.752	0.730	0.932	0.908	0.729	0.737	0.877	0.784	0.784	0.780	0.783	0.002
evn	grammar	2250	0.460	0.368	0.551	0.559	0.374	0.447	0.554	0.470	0.473	0.479	0.474	0.005
cni	grammar	105	0.992	0.991	0.999	0.993	0.996	0.997	0.993	0.993	0.996	0.995	0.995	0.002
hai	wiki	31	0.715	0.656	0.874	0.731	0.731	0.708	0.773	0.728	0.727	0.717	0.724	0.006
ntu	IGT	560	0.800	0.762	0.947	0.917	0.772	0.766	0.897	0.811	0.792	0.806	0.803	0.010

Table.3: Model accuracies for each sampling strategy, across 30 different languages. Data used to generate Figure.3.

Chapter 6

Conclusion

One of the goals of this thesis was to examine the use of computational resources in aid of language documentation. Data processing is often the primary use of computation in the context of language documentation. For example, many resources have been developed for the first part of the annotation process - transcription. Speech-to-text tools, predicated on both old and new computational methods, are used to alleviate the resource demands placed primarily on the linguist(s). The same constraints are observed in the next stage of annotation – glossing. Often this step requires more linguistic expertise. With the success of computational methods for transcription, the long-standing robustness of finite-state approaches to phonology and morphology, and now the success of neural-based models, it is no surprise that computational methods are recruited to reduce the strain on resources required. Although much of the architecture exists, the application to field linguistics requires refinement.

The contribution of this thesis is twofold; computational resources for the Papuan language Nen and a more general investigation of the interface between computational and field linguistic methods. These considerations include how much detail the gloss should include, what kind of data is best for model building, whether the data requirements of NLP models and linguistic corpora intersect, and whether data tailored to NLP models help reveal linguistic insight.

The first half of this thesis focused on building morphological models for Nen verbs. Both FST and neural network approaches and their corresponding considerations were presented. The first followed the traditional approach of FSTs. The motivation for adopting such an approach is clear, given the low-resource setting and ongoing documentation efforts. A rule-based method allows for quick adaptations when forms are discovered or description changes. Furthermore, it leverages heuristic linguistic insight that might not yet be substantiated in a statistically meaningful way via data. Two segmentation approaches are compared to further investigate architectural decisions made when building an FST: 'chunking'[1] and complete decomposition. In the 'chunking' approach, the suffix treated as one unit. Similarly, the prefix can be seen as a concatenation of the undergoer, the directional and future imperative prefix in the maximal case. By contrast, complete decomposition follows the linguistic description where both the prefix and suffix can be split further into the components noted in 1.3.

The results showed that despite the strong linguistic motivations for a morpheme-based approach, the computational implementation remains largely indifferent to the two approaches. The main difference is in the resulting FST size. While this may appear to be a secondary concern, it can become more pressing when a large lexicon is available or the language described displays complex morphophonology. In particular, this may become a primary consideration if the FST is to be used in-field where computational resources are typically limited.

Technological advances and the availability of computing power have led to neural-based approaches dominating NLP. This extends to the realm of phonology and morphology. Sequence-to-sequence models have been used to model morphological inflection. The SIGMORPHON share tasks have pioneered the inflection task and, over the years, have shown the massive success of neural-based architecture. One

[1] Retrospectively, this term might be confusing given its usage in syntax, but here it is intended to mean 'treating the suffix as one unit rather than breaking further into minimal components as expected of a morpheme'.

of the primary benefits of neural-based approaches is the ability to model complex phenomena without the need for domain expertise. It also can capture the statistical nature of language[2]. With these features in mind, the analysis of Nen verbal morphology is extended from FSTs to models based on deep learning architectures.

Two high-performing models from the 2017 SIGMORPHON shared task for low-resourced languages were compared. Given the statistical nature of these architectures, the effects of training data composition were investigated. First, the effects of how the sampling method by which the training data was constructed on model performance were examined. The training data was based on the corpus Zipfian distribution and compared with a random sampling strategy. The former approach is motivated by the Zipfian nature of language; word forms are not encountered uniformly, and often a handful of forms are encountered with higher frequency than the rest of the paradigm. A minor difference in model performance was observed, with random sampling marginally outperforming the Zipfian counterpart. This result was unsurprising as random sampling allows for a more diverse paradigm representation. The Zipfian distribution was calculated from the existing Nen corpus, and as with any corpus, the resultant distribution is conditioned mainly by the semantics of the recordings available. This line of inquiry was furthered by introducing training size as a variable. The experimental setup followed the 2017 SIGMORPHON shared task, namely 100 training samples (Low resource (LR)), 1,000 samples (Medium resource (MR)), and 10,000 (High resource (HR)). An additional 'ALL' setting was included to utilise the available corpus at 2,260 forms.

Second, the correlation between model performance per verb type and verb type representation in training was studied. As expected, when only one verb type is encountered, the model best predicts that verb type. However, when an egalitarian approach is adopted, whereby the model is exposed to an equal amount of each verb type, the model performance is not equal across verb types.

[2]FSTs can also model statistical behaviour with weighted FSTs, where in addition to the input/output pairs a weight is included.

Lastly, a special test case to probe whether each architecture can learn and predict paradigmatic structures was presented. The test case focuses on syncretism. Across the verbal paradigm, Nen exhibits syncretism between the second and third-person ('2|3sgA') singular cells. This pattern does not hold for the perfective past ('PFV.PST') TAM category. The second singular person ('2sgA') tag is with-held during training to test whether the syncretic behaviour is learnt. The model is exposed only to the third singular ('3sgA') forms. When the second singular person ('2sgA') form is elicited during test time, both models overwhelmingly (80% and 90% of the examples) predict syncretism by producing the third singular ('3sgA') form.

The second half of this thesis concentrated on utilising these computational constructions to address corpus concerns that are of interest to field linguists. One such question is the amount of data needed to capture the intricacies of a language. Using a similar setup as the neural approach to Nen verbal morphology, a SIGMORPHON neural baseline model is trained with various data quantities and a learning curve was generated. At test time, the model was asked to inflect full tables for a selected verb for each verb type. Inflected forms of the verb chosen for the test were excluded at training to avoid pre-exposure. During training, the infinitive forms of the test verbs were included to account for the performance dip noted for previously unseen lemma. The test verbs were chosen based on three factors: availability of tables verified by native Nen speakers, regularity of inflection patterns and similarity in phonology to the corpus-account counterpart. The accuracy for each verb type can be seen as a type of 'coverage'. In other words, how well the model can account for a complete verbal paradigm. The model-based 'coverage' is compared to a corpus-based account, where the trajectory of one representative verb for each verb type is followed and tallied across all the unique forms encountered in the corpus.

This study proposed a new way to measure paradigm 'coverage'. Predictably, it showed that a model-based account covers more of the paradigm than would be naturally encountered in a corpus. It found that the model can achieve high performance

for low-frequency MSDs and provided a rough estimation of the data requirements to achieve full attestation of the paradigm.

Given the difference between the needs of a linguistic corpus and datasets for model training, the last part of this thesis examines the use of active learning to guide data collection, with particular attention to model performance. Four sampling strategies are compared, with model performance improving in the following method order: incorrect, low-model confidence, high entropy, and random. By including the opposite of these sampling strategies, the results showed that adding data without regard to sample diversity leads to overfitting and reduced model accuracy.

The work presented in this thesis is intended as a potential linguist guide for adopting NLP into the documentation process, detailing modelling decisions and their consequences (if any). Concentrating on Nen has made it possible to provide a pedagogical overview of computational resource development for morphological modelling. The interface between linguistic and computational practises was examined at each stage to outline where they converge and diverge. The purposes of these computational models are to aid analysis in some form by organising the data to draw focus to where linguistic insight is needed.

For example, both the FST and the neural-based models can be employed directly to generate inflectional tables of all verbs identified in the dictionary. At this point, the linguist or native speaker can verify the predicted forms. An additional step might be to cross-reference the inflection tables produced to what is and is not attested in the corpus to better serve a documentation process. These tables can be used to guide linguist attention and data collection for archival purposes.

On the other hand, this is an area where computational considerations can feed into linguistic practises. AL was shown to be a useful tool for significantly improving model performance. Future work could investigate whether there are linguistic foundations to the model-identified low-confidence forms and how these insights can potentially interface with the documentation process. To this end, it is imperative

that an AL cycle is tested in a real-time documentation setting to substantiate its effectiveness in practice.

One oversight in implementing computational methods for documentation efforts is the assumption of a standardised way of working amongst fieldworkers. It is an understandable notion given the generalising power afforded by computers. To better understand the underpinning processes of field linguistics, gathering as many perspectives and insights into workflows as possible is essential. This might be addressed as extensive and, importantly, ongoing surveys and interviews with field linguists to determine what could be met by computational techniques. Beyond that, it would be incredibly insightful for a computational linguist to engage in fieldwork to compare perspectives and workflow and contextualise the research questions.

Another exciting extension to the work presented here would be the inclusion of a principal parts account to the question of how much data is needed to capture a linguistic sub-system. Given the statistical nature of neural networks, what does capturing the principal parts of a system entail? Is it possible to capture morphophonological patterns in highly frequent MSDs and extend the pattern to rare MSDs? All these research questions can potentially strengthen the interaction between the fields of computer science (specifically NLP) and linguistics. Given that the employment of computational methods in linguistics is becoming more commonplace, fostering a mutual symbiotic relationship is essential.

Bibliography

Ackerman, Farrell, James P. Blevins and Robert Malouf (July 2009). 'Parts and Wholes: Implicative Patterns in Inflectional Paradigms'. In: *Analogy in Grammar: Form and Acquisition*. ISBN: 9780199547548. DOI: 10.1093/acprof:oso/9780199547548.003.0003. eprint: https://academic.oup.com/book/0/chapter/150182758/chapter-ag-pdf/44982748/book_6411_section_150182758.ag.pdf. URL: https://doi.org/10.1093/acprof:oso/9780199547548.003.0003.

Ackerman, Farrell and Robert Malouf (2013). 'Morphological organization: The low conditional entropy conjecture'. In: *Language*, pp. 429–464. URL: https://www.jstor.org/stable/24671935.

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird and Alexis Michaud (2018). 'Evaluating phonemic transcription of low-resource tonal languages for language documentation'. In: *LREC 2018 (Language Resources and Evaluation Conference)*, pp. 3356–3365. URL: https://aclanthology.org/L18-1530.

Aharoni, Roee and Yoav Goldberg (July 2017). 'Morphological Inflection Generation with Hard Monotonic Attention'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada, pp. 2004–2015. DOI: 10.18653/v1/P17-1183.

Ahmad, Wasi, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang and Nanyun Peng (June 2019). 'On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2440–2452.

DOI: 10.18653/v1/N19-1253. URL: https://www.aclweb.org/anthology/N19-1253.

Anastasopoulos, Antonios, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto and David Chiang (Aug. 2018). 'Part-of-Speech Tagging on an Endangered Language: A Parallel Griko-Italian Resource'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2529–2539. URL: https://aclanthology.org/C18-1214.

Andriyanets, Vasilisa and Francis Tyers (Aug. 2018). 'A prototype finite-state morphological analyser for Chukchi'. In: *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 31–40. URL: https://www.aclweb.org/anthology/W18-4804.

Anik, Ariful Islam and Andrea Bunt (2021). 'Data-centric explanations: Explaining training data of machine learning systems to promote transparency'. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. URL: https://doi.org/10.1145/3411764.3445736.

Atkins, Sue, Jeremy Clear and Nicholas Ostler (Jan. 1992). 'Corpus Design Criteria'. In: *Literary and Linguistic Computing* 7.1, pp. 1–16. ISSN: 0268-1145. DOI: 10.1093/llc/7.1.1. eprint: https://academic.oup.com/dsh/article-pdf/7/1/1/10886900/1.pdf. URL: https://doi.org/10.1093/llc/7.1.1.

Austin, Peter (2005). 'Training for language documentation: The SOAS experience'. In: *Linguistics Society of America Conference on Language Documentation: Theory, Practice, and Values. July*, pp. 9–11. URL: http://www.ddl.ish-lyon.cnrs.fr/fulltext/Grinevald/SOCIOLINGUISTIQUE_M1/austin/Handout%20P.Austin%20Lyon%2028-03-07.pdf.

Austin, Peter and Julia Sallabank (2011). *The Cambridge handbook of endangered languages*. URL: https://doi.org/10.1017/CBO9780511975981.

Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio (2015). 'Neural machine translation by jointly learning to align and translate'. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*. URL: http://arxiv.org/abs/1409.0473.

Baird, Louise, Nicholas Evans and Simon J. Greenhill (2022). 'Blowing in the wind: Using 'North Wind and the Sun' texts to sample phoneme inventories'. In: *Journal of the International Phonetic Association* 52.3, pp. 453–494. DOI: 10.1017/S002510032000033X.

Baldridge, Jason and Alexis Palmer (Aug. 2009). 'How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation.' In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 296–305. URL: https://aclanthology.org/D09-1031.

Banko, Michele and Eric Brill (2001). 'Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing'. In: *Proceedings of the first international conference on Human language technology research*. URL: https://aclanthology.org/H01-1052.

Barriga Martínez, Diego, Victor Mijangos and Ximena Gutierrez-Vasques (June 2021). 'Automatic Interlinear Glossing for Otomi language'. In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, pp. 34–43. DOI: 10.18653/v1/2021.americasnlp-1.5. URL: https://aclanthology.org/2021.americasnlp-1.5.

Barth, Danielle and Stefan Schnell (2021). *Understanding corpus linguistics*. Routledge. URL: https://doi.org/10.4324/9780429269035.

Batsuren, Khuyagbaatar, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell and Ekaterina Vylomova (July 2022a). 'The SIGMORPHON 2022 Shared Task on Morpheme Segmentation'. In: *Proceedings of the 19th SIGMORPHON Workshop on Computational*

Research in Phonetics, Phonology, and Morphology. Seattle, Washington: Association for Computational Linguistics, pp. 103–116. DOI: 10.18653/v1/2022.sigmorphon-1.11. URL: https://aclanthology.org/2022.sigmorphon-1.11.

Batsuren, Khuyagbaatar et al. (June 2022b). 'UniMorph 4.0: Universal Morphology'. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 840–855. URL: https://aclanthology.org/2022.lrec-1.89.

Beemer, Sarah, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, Piyush Mishra, Saliha Muradoglu, Lan Sang, Tyler Short, Sagarika Shreevastava, Elizabeth Spaulding, Testumichi Umada, Beilei Xiang, Changbing Yang and Mans Hulden (July 2020). 'Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars'. In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics, pp. 162–170. DOI: 10.18653/v1/2020.sigmorphon-1.18. URL: https://aclanthology.org/2020.sigmorphon-1.18.

Beesley, Kenneth R. and Lauri Karttunen (2003). *Finite-state morphology: Xerox tools and techniques*. CSLI, Stanford. ISBN: 1575864347.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad and James Glass (July 2017). 'What do Neural Machine Translation Models Learn about Morphology?' In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 861–872. DOI: 10.18653/v1/P17-1080. URL: https://aclanthology.org/P17-1080.

Bender, Emily M (2011). 'On achieving and evaluating language-independence in NLP'. In: *Linguistic Issues in Language Technology* 6. URL: https://journals.colorado.edu/index.php/lilt/article/download/1239/1077.

Berko, Jean (1958). 'The Child's Learning of English Morphology'. In: *WORD* 14.2-3, pp. 150–177. DOI: 10.1080/00437956.1958.11659661.

Bianco, Joseph Lo (2002). 'Real world language politics and policy'. In: *Language policy: Lessons from global models. Monterey, California: Monterey Institute of International Studies*.

Biber, Douglas (Jan. 1990). 'Methodological issues regarding corpus-based analyses of linguistic variation'. In: *Literary and Linguistic Computing* 5.4, pp. 257–269. ISSN: 0268-1145. DOI: 10.1093/llc/5.4.257. eprint: https://academic.oup.com/dsh/article-pdf/5/4/257/10889590/257.pdf. URL: https://doi.org/10.1093/llc/5.4.257.

Biber, Douglas (1993). 'Representativeness in corpus design'. In: *Literary and linguistic computing* 8.4, pp. 243–257. URL: https://doi.org/10.1093/llc/8.4.243.

Biber, Douglas, Susan Conrad and Randi Reppen (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511804489.

Bird, Steven (Nov. 2015). 'Email'. In: *Resource Network for Linguistic Diversity Discussion List*.

Bird, Steven (May 2022). 'Local Languages, Third Spaces, and other High-Resource Scenarios'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 7817–7829. DOI: 10.18653/v1/2022.acl-long.539. URL: https://aclanthology.org/2022.acl-long.539.

Black, H Andrew and Gary F Simons (2006). 'The SIL Field-Works Language Explorer approach to morphological parsing'. In: *Computational Linguistics for Lessstudied Languages: Texas Linguistics Society* 10. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9a82c018ab9b3d162103cf0cf4174e02ea7a7ed6.

Blevins, James P, Petar Milin and Michael Ramscar (2017). 'The Zipfian paradigm cell filling problem'. In: *Perspectives on morphological organization: Data and analyses* 10, p. 141. URL: https://doi.org/10.1163/9789004342934_008.

Blevins, James P. and Juliette Blevins (2009). *Analogy in grammar: Form and acquisition*. Oxford University Press. URL: https://doi.org/10.1093/acprof:oso/9780199547548.001.0001.

Boersma, Paul and David Weenink (2018). 'Praat: Doing phonetics by computer [Computer program]. Version 6.0. 37'. In: URL: http://www.praat.org/.

Bostrom, Kaj and Greg Durrett (Nov. 2020). 'Byte Pair Encoding is Suboptimal for Language Model Pretraining'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4617–4624. DOI: 10.18653/v1/2020.findings-emnlp.414. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.414.

Brezina, Vaclav (2018). Cambridge University Press, pp. 38–65. DOI: https://doi.org/10.1017/9781316410899.

Bybee, Joan L (1991). 'Natural morphology: The organization of paradigms and language acquisition'. In: *Crosscurrents in second language acquisition and linguistic theories* 2, pp. 67–92. DOI: https://doi.org/10.1075/lald.2.08byb.

Carroll, Matthew J. (2016). 'The Ngkolmpu Language'. PhD thesis. The Australian National University. DOI: https://doi.org/10.25911/5d74e0cfd5b85.

Ćavar, Damir, Malgorzata Cavar and Lwin Moe (2016). 'Global Open Resources and Information for Language and Linguistic Analysis (GORILLA)'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4484–4491. URL: https://aclanthology.org/L16-1710/.

Celikyilmaz, Asli, Elizabeth Clark and Jianfeng Gao (2020). 'Evaluation of Text Generation: A Survey'. In: *ArXiv* abs/2006.14799. URL: https://doi.org/10.48550/arXiv.2006.14799.

Chan, Erwin and Charles D Yang (2008). 'Structures and distributions in morphology learning'. PhD thesis. University of Pennsylvania. URL: https://www.proquest.

com/openview/1b87217dff0cbeb1ab330e5a8c83c9bf/1?pq-origsite=gscholar&cbl=18750.

Chandra, Vikram (n.d.). *Pāṇini: Catching the Ocean in a Cow's Hoofprint*. URL: http://hipporeads.com/pa%E1%B9%87ini-catching-the-ocean-in-a-cows-hoofprint/.

Chen, Emily and Lane Schwartz (2018). 'A morphological analyzer for St. Lawrence Island/Central Siberian Yupik'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. URL: https://aclanthology.org/L18-1416.

Chomsky, Noam (1956). 'Three models for the description of language'. In: *IRE Transactions on Information Theory* 2.3, pp. 113–124. DOI: 10.1109/TIT.1956.1056813.

Chomsky, Noam (1959). 'On certain formal properties of grammars'. In: *Information and Control* 2.2, pp. 137–167. ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(59)90362-6. URL: https://www.sciencedirect.com/science/article/pii/S0019995859903626.

Çöltekin, Çagri (2014). 'A set of open source tools for Turkish natural language processing.' In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 1079–1086. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/437_Paper.pdf.

Comrie, Bernard, Martin Haspelmath and Balthasar Bickel (2008). 'The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses'. In: *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig.*, Last updated 2015. URL: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.

Cotterell, Ryan, Christo Kirov, Mans Hulden and Jason Eisner (Mar. 2019a). 'On the Complexity and Typology of Inflectional Morphological Systems'. In: *Transactions of the Association for Computational Linguistics* 7, pp. 327–342. DOI: 10.1162/tacl_a_00271. URL: https://www.aclweb.org/anthology/Q19-1021.

Cotterell, Ryan, Christo Kirov, Mans Hulden and Jason Eisner (June 2019b). 'On the Complexity and Typology of Inflectional Morphological Systems'. In: *Transactions of the Association for Computational Linguistics* 7, pp. 327–342. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00271. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00271/1923163/tacl_a_00271.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00271.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner and Mans Hulden (Nov. 2018). 'The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection'. In: *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*. Ed. by Mans Hulden and Ryan Cotterell. Brussels: Association for Computational Linguistics, pp. 1–27. DOI: 10.18653/v1/K18-3001. URL: https://aclanthology.org/K18-3001.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner and Mans Hulden (Aug. 2017). 'CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages'. In: *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Vancouver: Association for Computational Linguistics, pp. 1–30. DOI: 10.18653/v1/K17-2001. URL: https://www.aclweb.org/anthology/K17-2001.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner and Mans Hulden (Aug. 2016). 'The SIGMORPHON 2016 Shared Task—Morphological Reinflection'. In: *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Berlin, Germany: Association for Computational Linguistics, pp. 10–22. DOI: 10.18653/v1/W16-2002. URL: https://www.aclweb.org/anthology/W16-2002.

Crystal, David (2002). *Language Death*. Canto. DOI: 10.1017/CBO9781139871549.

De Mulder, Wim, Steven Bethard and Marie-Francine Moens (2015). 'A survey on the application of recurrent neural networks to statistical language modeling'. In: *Computer Speech & Language* 30.1, pp. 61–98. URL: https://doi.org/10.1016/j.csl.2014.09.005.

Dehouck, Mathieu and Pascal Denis (Oct. 2018). 'A Framework for Understanding the Role of Morphology in Universal Dependency Parsing'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2864–2870. DOI: 10.18653/v1/D18-1312. URL: https://aclanthology.org/D18-1312.

Dreyfus, Tommy (1991). 'Advanced Mathematical Thinking Processes'. In: *Advanced Mathematical Thinking*. Ed. by David Tall. Dordrecht: Springer Netherlands, pp. 25–41. ISBN: 978-0-306-47203-9. DOI: 10.1007/0-306-47203-1_2. URL: https://doi.org/10.1007/0-306-47203-1_2.

Durantin, Gautier, Ben Foley, Nicholas Evans and Janet Wiles (2017). *Transcription survey*.

Eskander, Ramy, Francesca Callejas, Elizabeth Nichols, Judith Klavans and Smaranda Muresan (May 2020). 'MorphAGram, Evaluation and Framework for Unsupervised Morphological Segmentation'. English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 7112–7122. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.879.

Eskander, Ramy, Judith Klavans and Smaranda Muresan (Aug. 2019). 'Unsupervised Morphological Segmentation for Low-Resource Polysynthetic Languages'. In: *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Florence, Italy: Association for Computational Linguistics, pp. 189–195. DOI: 10.18653/v1/W19-4222. URL: https://aclanthology.org/W19-4222.

Evans, Nicholas (2008). 'Review of Gippert, Jost, Nikolaus Himmelmann and Ulrike Mosel (eds.), Essentials of language documentation'. In: *Language Documentation & Conservation* 2, pp. 340–350. URL: https://scholarspace.manoa.hawaii.edu/

server/api/core/bitstreams/b437d9c5-f768-4cb5-873f-6fa61e39e40e/content.

Evans, Nicholas (2012). 'Even more diverse than we had thought: The multiplicity of Trans-Fly languages'. In: *Melanesian languages on the edge of Asia: Challenges for the 21st century*. University of Hawaii Press. URL: http://hdl.handle.net/1885/28667.

Evans, Nicholas (2014). 'Positional verbs in Nen'. In: *Oceanic Linguistics* 53.2, pp. 225–255. ISSN: 1527-9421. DOI: http://doi.org/10.1353/ol.2014.0019.

Evans, Nicholas (2015). 'Chapter 26: Valency in Nen'. In: *Volume 2 Case Studies from Austronesia, the Pacific, the Americas, and Theoretical Outlook*. Ed. by Andrej Malchukov and Bernard Comrie. Berlin, München, Boston: De Gruyter Mouton, pp. 1069–1116. ISBN: 9783110429343. DOI: doi:10.1515/9783110429343-006. URL: https://doi.org/10.1515/9783110429343-006.

Evans, Nicholas (2016). 'Inflection in Nen'. In: *The Oxford Handbook of Inflection*. Ed. by Matthew Baerman. Oxford University Press, USA, pages 543–575. DOI: 10.1093/oxfordhb/9780199591428.013.23.

Evans, Nicholas (2017). 'Quantification in Nen'. In: *Handbook of Quantifiers in Natural Language: Volume II*. Cham: Springer International Publishing, pp. 571–607. ISBN: 978-3-319-44330-0. DOI: 10.1007/978-3-319-44330-0_11. URL: https://doi.org/10.1007/978-3-319-44330-0_11.

Evans, Nicholas (2019a). 'Nen dictionary'. In: *Dictionaria*, pp. 1–5005. URL: https://dictionaria.clld.org/contributions/nen.

Evans, Nicholas (2019b). 'Waiting for the Word: Distributed Deponency and the Semantic Interpretation of Number in the Nen Verb'. In: *Morphological Perspectives. Papers In Honour of Greville G. Corbett*, pp. 100–123. DOI: doi:10.1515/9781474446020-006. URL: https://doi.org/10.1515/9781474446020-006.

Evans, Nicholas (2020). 'One thousand and one coconuts: Growing memories in Southern New Guinea'. In: *The Contemporary Pacific* 32.1, pp. 72–96. URL: https://www.cabidigitallibrary.org/doi/full/10.5555/20220369286.

Evans, Nicholas (2022). *Words of wonder: Endangered languages and what they tell us*. John Wiley & Sons. URL: https://www.wiley.com/en-us/9781119758754.

Evans, Nicholas (n.d.). 'Grammar of Nen'.

Evans, Nicholas and Stephen C. Levinson (2009). 'The myth of language universals: Language diversity and its importance for cognitive science'. In: *Behavioral and Brain Sciences* 32.5, pp. 429–448. DOI: 10.1017/S0140525X0999094X.

Evans, Nicholas and Julia Colleen Miller (2016). 'Nen'. In: *Journal of the International Phonetic Association* 46.3, pp. 331–349. URL: https://doi.org/10.1017/S0025100315000365.

Evans, Nicholas and Hans-Jürgen Sasse (2002). 'Introduction: Problems of polysynthesis'. In: *Problems of Polysynthesis*. Ed. by Nicholas Evans and Hans-Jürgen Sasse. Berlin: Akademie Verlag, pp. 1–14. ISBN: 9783050080956. DOI: doi:10.1524/9783050080956.1. URL: https://doi.org/10.1524/9783050080956.1.

Evans, Nicholas and Hans-Jurgen Sasse (2007). 'Searching for meaning in the Library of Babel: Field semantics and problems of digital archiving'. English. In: *Archives and Social Studies: A Journal of Interdisciplinary Research* 1, pp. 63–123. URL: https://archivo.cartagena.es/doc/Archivos_Social_Studies/Vol1_n0/05-evanssasse_searching.pdf.

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

Faraoni, Valerio (Aug. 2020). 'Natural phenomena described by the same equation'. In: *European Journal of Physics* 41.5, p. 054002. DOI: 10.1088/1361-6404/ab9c93. URL: https://dx.doi.org/10.1088/1361-6404/ab9c93.

Faruqui, Manaal, Yulia Tsvetkov, Graham Neubig and Chris Dyer (June 2016). 'Morphological Inflection Generation Using Character Sequence to Sequence Learning'. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 634–643. DOI: 10.18653/v1/N16-1077. URL: https://www.aclweb.org/anthology/N16-1077.

Finkel, Raphael and Gregory Stump (2007). 'Principal parts and morphological typology'. In: *Morphology* 17.1, pp. 39–75. URL: https://doi.org/10.1007/s11525-007-9115-9.

Foley, Ben, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash et al. (2018). 'Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS).' In: *SLTU*, pp. 205–209. URL: https://doi.org/10.21437/sltu.2018-43.

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart and Anna Korhonen (Oct. 2018). 'On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 316–327. DOI: 10.18653/v1/D18-1029. URL: https://aclanthology.org/D18-1029.

Gil, David (2001). 'Escaping Eurocentrism: Fieldwork as a process of unlearning'. In: *Linguistic fieldwork*, pp. 102–132. URL: https://doi.org/10.1017/CBO9780511810206.006.

Ginn, Michael, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden and Miikka Silfverberg (July 2023). 'Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing'. In: *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Toronto, Canada: Association for Computational Linguistics, pp. 186–201. URL: https://aclanthology.org/2023.sigmorphon-1.20.

Goldman, Omer, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty and Ekaterina Vylomova (July 2023). 'SIGMORPHON UniMorph 2023 Shared Task 0: Typologically Diverse Morphological Inflection'. In: *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Toronto, Canada: Association for Computational

Linguistics, pp. 117–125. URL: https://aclanthology.org/2023.sigmorphon-1.13.

Goldsmith, John (2001). 'Unsupervised Learning of the Morphology of a Natural Language'. In: *Computational Linguistics* 27.2, pp. 153–198. DOI: 10.1162/089120101750300490. URL: https://aclanthology.org/J01-2001.

Gorman, Kyle, Arya McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg and Magdalena Markowska (Nov. 2019). 'Weird Inflects but OK: Making Sense of Morphological Generation Errors'. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 140–151. DOI: 10.18653/v1/K19-1014. URL: https://aclanthology.org/K19-1014.

Gorman, Kyle and Richard Sproat (2016). 'Minimally Supervised Number Normalization'. In: *Transactions of the Association for Computational Linguistics* 4, pp. 507–519. DOI: 10.1162/tacl_a_00114. URL: https://www.aclweb.org/anthology/Q16-1036.

Goyal, Pawan, Amba Kulkarni and Laxmidhar Behera (2007). 'Computer simulation of aṣṭādhyāyı: Some insights'. In: *Sanskrit Computational Linguistics*, pp. 139–161. URL: https://doi.org/10.1007/978-3-642-00155-0_5.

Greenberg, Joseph H (1960). 'A quantitative approach to the morphological typology of language'. In: *International journal of American linguistics* 26.3, pp. 178–194. URL: https://doi.org/10.1086/464575.

Gries, Stefan Th (2008). 'Dispersions and Adjusted Frequencies in Corpora'. In: *International Journal of Corpus Linguistics* 13.4, pp. 403–437. DOI: https://doi.org/10.1075/ijcl.13.4.02gri.

Gupta, Nitin, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal and Vitobha Munigala (2021). 'Data Quality for Machine Learning Tasks'. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing

Machinery, pp. 4040–4041. ISBN: 9781450383325. DOI: 10.1145/3447548.3470817. URL: https://doi.org/10.1145/3447548.3470817.

Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl and Alexandra Birch (Sept. 2022). 'Survey of Low-Resource Machine Translation'. In: *Computational Linguistics* 48.3, pp. 673–732. ISSN: 0891-2017. DOI: 10.1162/coli_a_00446. URL: https://doi.org/10.1162/coli%5C_a%5C_00446.

Harris, Alice C (2017). *Multiple exponence*. Oxford University Press. DOI: 10.1093/acprof:oso/9780190464356.001.0001.

Haspelmath, Martin, Matthew S Dryer, David Gil and Bernard Comrie (2005). *The world atlas of language structures*. OUP Oxford.

Henderson, Peter, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup and David Meger (2018). 'Deep reinforcement learning that matters'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. URL: https://doi.org/10.1609/aaai.v32i1.11694.

Himmelmann, Nikolaus P (1998). 'Documentary and descriptive linguistics'. In: URL: https://doi.org/10.1515/ling.1998.36.1.161.

Himmelmann, Nikolaus P et al. (2006). 'Language documentation: What is it and what is it good for'. In: *Essentials of language documentation* 178.1. URL: https://doi.org/10.1515/9783110197730.1.

Himmelmann, Nikolaus P (2018). 'Meeting the transcription challenge'. In: URL: http://hdl.handle.net/10125/24806.

Hovy, Eduard and Julia Lavid (2010). 'Towards a 'science'of corpus annotation: A new methodological challenge for corpus linguistics'. In: *International journal of translation* 22.1, pp. 13–36. URL: https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat and Melvin Johnson (July 2020). 'XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation'. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh.

Vol. 119. Proceedings of Machine Learning Research. Virtual: PMLR, pp. 4411–4421. URL: http://proceedings.mlr.press/v119/hu20b.html.

Hulden, Mans (Apr. 2009). 'Foma: A Finite-State Compiler and Library'. In: *Proceedings of the Demonstrations Session at EACL 2009*. Athens, Greece: Association for Computational Linguistics, pp. 29–32. URL: https://www.aclweb.org/anthology/E09-2008.

Hulden, Mans (June 2022). 'Finite-State Technology'. In: *The Oxford Handbook of Computational Linguistics*. ISBN: 9780199573691. DOI: 10.1093/oxfordhb/9780199573691.013.39. eprint: https://academic.oup.com/book/0/chapter/358149188/chapter-pdf/45719713/oxfordhb-9780199573691-e-39.pdf. URL: https://doi.org/10.1093/oxfordhb/9780199573691.013.39.

Hyman, Larry M. (2007). 'Elicitation as Experimental Phonology'. In: *Experimental Approaches to Phonology*, pp. 7–24. URL: https://escholarship.org/content/qt4nr5n1z0/qt4nr5n1z0_noSplash_a490fdbb1a7c22d2b95bd52d089dbfb8.pdf?t=pdtko3.

Ide, Nancy and James Pustejovsky, eds. (2017). *Handbook of Linguistic Annotation*. 1st. Springer Publishing Company, Incorporated. ISBN: 9402408797. URL: https://doi.org/10.1007/978-94-024-0881-2.

Iqbal, Touseef and Shaima Qureshi (2022). 'The survey: Text generation models in deep learning'. In: *Journal of King Saud University - Computer and Information Sciences* 34.6, Part A, pp. 2515–2528. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2020.04.001. URL: https://www.sciencedirect.com/science/article/pii/S1319157820303360.

Johnson, C Douglas (1972). 'Formal Aspects of Phonological Description.(1970 doctoral dissertation, UC Berkeley.) Mouton & Co'. In: *The Hague*. URL: https://pages.ucsd.edu/~ebakovic/compphon/Johnson%201972%201-up.pdf.

Johnson, Mark, Peter Anderson, Mark Dras and Mark Steedman (2018). 'Predicting accuracy on large datasets from smaller pilot data'. In: *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2, pp. 450–455. URL: https://aclanthology.org/P18-2072.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali and Monojit Choudhury (July 2020). 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. URL: https://aclanthology.org/2020.acl-main.560.

Jurafsky, Dan and James H. Martin (2009). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall. URL: http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.

Kann, Katharina and Hinrich Schütze (Aug. 2016). 'Single-Model Encoder-Decoder with Explicit Morphological Representation for Reinflection'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 555–560. DOI: 10.18653/v1/P16-2090. URL: https://www.aclweb.org/anthology/P16-2090.

Kaplan, Ronald M. and Martin Kay (1994). 'Regular Models of Phonological Rule Systems'. In: *Computational Linguistics* 20.3, pp. 331–378. URL: https://aclanthology.org/J94-3001.

Kazeminejad, Ghazaleh, Andrew Cowell and Mans Hulden (2017). 'Creating lexical resources for polysynthetic languages—the case of Arapaho'. In: *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 10–18. URL: https://aclanthology.org/W17-0102.

Kelly, Barbara, Gillian Wigglesworth, Rachel Nordlinger and Joseph Blythe (2014). 'The Acquisition of Polysynthetic Languages'. In: *Language and Linguistics Compass* 8.2, pp. 51–64. DOI: https://doi.org/10.1111/lnc3.12062. eprint: https:

//compass.onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12062. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12062.

Khudanpur, Sanjeev P (2006). 'Multilingual language modeling'. In: *Multilingual Speech Processing*, p. 169.

Kilgarriff, Adam (June 1997). 'Putting frequencies in the dictionary'. In: *International Journal of Lexicography* 10.2, pp. 135–155. ISSN: 0950-3846. DOI: 10.1093/ijl/10.2.135. eprint: https://academic.oup.com/ijl/article-pdf/10/2/135/9820144/135.pdf. URL: https://doi.org/10.1093/ijl/10.2.135.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner and Mans Hulden (May 2018a). 'UniMorph 2.0: Universal Morphology'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1293.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J Mielke, Arya McCarthy, Sandra Kübler et al. (2018b). 'UniMorph 2.0: Universal Morphology'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. URL: https://aclanthology.org/L18-1293.

Klein, Stav and Reut Tsarfaty (July 2020). 'Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology?' In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics, pp. 204–209. DOI: 10.18653/v1/2020.sigmorphon-1.24. URL: https://aclanthology.org/2020.sigmorphon-1.24.

Kodner, Jordan, 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 (July 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*. Ed. by Garrett Nicolai and Eleanor Chodroff. Seattle, Washington: Association for Computational Linguistics, pp. 176–203. DOI: 10.18653/v1/2022.sigmorphon-1.19. URL: https://aclanthology.org/2022.sigmorphon-1.19.

Kolmogorov, Andrei N (1963). 'On tables of random numbers'. In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 369–376. URL: http://www.jstor.org/stable/25049284.

Koskenniemi, Kimmo (1983). *Two-level morphology: A general computational model for word-form recognition and production*. Vol. 11. University of Helsinki, Department of General Linguistics Helsinki, Finland. URL: https://aclanthology.org/P84-1038.

Krauss, Michael (1992). 'The world's languages in crisis'. In: *Language* 68.1, pp. 4–10.

Kruskal, William and Frederick Mosteller (1980). 'Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939'. In: *International Statistical Review / Revue Internationale de Statistique* 48.2, pp. 169–195. ISSN: 03067734, 17515823. URL: http://www.jstor.org/stable/1403151 (visited on 30/07/2023).

Lachler, Jordan, Lene Antonsen, Trond Trosterud, Sjur Moshagen and Antti Arppe (2018). 'Modeling Northern Haida Verb Morphology'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. URL: https://aclanthology.org/L18-1368.

Lane, William and Steven Bird (2019). 'Towards A Robust Morphological Analyzer for Kunwinjku'. In: *Proceedings of the The 17th Annual Workshop of the Australasian*

Language Technology Association, pp. 1–9. URL: https://aclanthology.org/U19-1001.

Langendoen, D. Terence (1981). 'The Generative Capacity of Word-Formation Components'. In: *Linguistic Inquiry* 12.2, pp. 320–322. ISSN: 00243892, 15309150. URL: http://www.jstor.org/stable/4178223.

Le Ferrand, Éric, Steven Bird and Laurent Besacier (Oct. 2022). 'Fashioning Local Designs from Generic Speech Technologies in an Australian Aboriginal Community'. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4274–4285. URL: https://aclanthology.org/2022.coling-1.376.

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (2015). 'Deep learning'. In: *Nature* 521.7553, pp. 436–444. URL: https://doi.org/10.1038/nature14539.

Leech, Geoffrey (1991). 'The State of the Art in Corpus Linguistics'. In: *English Corpus Linguistics*, pp. 20–41. DOI: https://doi.org/10.4324/9781315845890-11.

Lehmann, Christian (2001). *Language documentation: A program*. na. URL: https://www.christianlehmann.eu/publ/Language_documentation.pdf.

Lignos, Constantine, Charles D. Yang, Andrew Hippisley and Gregory T. Stump (2016). 'Morphology and Language Acquisition'. In: URL: https://doi.org/10.1002/9781405166348.ch19.

Lindén, Krister, Erik Axelson, Sam Hardwick, Tommi A Pirinen and Miikka Silfverberg (2011). 'Hfst—framework for compiling and applying morphologies'. In: *International Workshop on Systems and Frameworks for Computational Morphology*. Springer, pp. 67–85. URL: https://doi.org/10.1007/978-3-642-23138-4_5.

Lipton, Zachary C and Jacob Steinhardt (2019). 'Research for practice: Troubling trends in machine-learning scholarship'. In: *Communications of the ACM* 62.6, pp. 45–53. URL: https://doi.org/10.1145/3316774.

Liu, Ling (2021). 'Morphological Generation with Deep Learning Approaches'. PhD thesis. University of Colorado. URL: https://www.proquest.com/openview/

22f913db2b4f7626f87578111abfb337/1?pq-origsite=gscholar&cbl=18750&diss=y.

Liu, Ling and Mans Hulden (July 2020). 'Leveraging Principal Parts for Morphological Inflection'. In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics, pp. 153–161. DOI: 10.18653/v1/2020.sigmorphon-1.17. URL: https://aclanthology.org/2020.sigmorphon-1.17.

Liu, Ling and Mans Hulden (May 2022). 'Detecting Annotation Errors in Morphological Data with the Transformer'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 166–174. URL: https://aclanthology.org/2022.acl-short.19.

Lloret, Elena and Manuel Palomar (2012). 'Text summarisation in progress: A literature review'. In: *Artificial Intelligence Review* 37, pp. 1–41. URL: https://doi.org/10.1007/s10462-011-9216-z.

Luong, Thang, Hieu Pham and Christopher D. Manning (Sept. 2015). 'Effective Approaches to Attention-based Neural Machine Translation'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421. DOI: 10.18653/v1/D15-1166. URL: https://aclanthology.org/D15-1166.

Lüpke, Friederike (2009). 'Data collection methods for field-based language documentation'. In: *Language documentation and description* 6, pp. 53–100. URL: https://eprints.soas.ac.uk/id/eprint/12420.

Macklin-Cordes, Jayden L and Erich R Round (2020). 'Re-evaluating phoneme frequencies'. In: *Frontiers in psychology* 11, p. 570895. URL: https://doi.org/10.3389/fpsyg.2020.570895.

Makarov, Peter and Simon Clematide (Oct. 2018a). 'Imitation Learning for Neural Morphological String Transduction'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for

Computational Linguistics, pp. 2877–2882. DOI: 10.18653/v1/D18-1314. URL: https://aclanthology.org/D18-1314.

Makarov, Peter and Simon Clematide (Aug. 2018b). 'Neural Transition-based String Transduction for Limited-Resource Setting in Morphology'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 83–93. URL: https://www.aclweb.org/anthology/C18-1008.

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre and Daniel Zeman (June 2021). 'Universal Dependencies'. In: *Computational Linguistics* 47.2, pp. 255–308. DOI: 10.1162/coli_a_00402. URL: https://aclanthology.org/2021.cl-2.11.

Mathews, Peter H (1974). *Morphology: An introduction to the theory of word-structure*. Cambridge, England: Cambridge University Press. DOI: 10.1017/S0022226700004588.

McCarthy, Arya, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden and David Yarowsky (May 2020). 'UniMorph 3.0: Universal Morphology'. English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 3922–3931. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.483.

McCarthy, Arya, Miikka Silfverberg, Ryan Cotterell, Mans Hulden and David Yarowsky (Nov. 2018). 'Marrying Universal Dependencies and Universal Morphology'. In: *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Ed. by

Marie-Catherine de Marneffe, Teresa Lynn and Sebastian Schuster. Brussels, Belgium: Association for Computational Linguistics, pp. 91–101. DOI: 10.18653/v1/W18-6011. URL: https://aclanthology.org/W18-6011.

McCarthy, Arya, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell and Mans Hulden (Aug. 2019). 'The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection'. In: *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Ed. by Garrett Nicolai and Ryan Cotterell. Florence, Italy: Association for Computational Linguistics, pp. 229–244. DOI: 10.18653/v1/W19-4226. URL: https://aclanthology.org/W19-4226.

McEnery, Tony, Richard Xiao and Yukio Tono (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis. URL: http://www.routledge.com/textbooks/0415286239/default.asp.

McMillan-Major, Angelina (Jan. 2020). 'Automating Gloss Generation in Interlinear Glossed Text'. In: *Proceedings of the Society for Computation in Linguistics 2020*. New York, New York: Association for Computational Linguistics, pp. 355–366. URL: https://aclanthology.org/2020.scil-1.42.

Mielke, Sabrina J., Ryan Cotterell, Kyle Gorman, Brian Roark and Jason Eisner (July 2019). 'What Kind of Language Is Hard to Language-Model?' In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4975–4989. DOI: 10.18653/v1/P19-1491. URL: https://www.aclweb.org/anthology/P19-1491.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013). 'Distributed Representations of Words and Phrases and Their Compositionality'. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger. Lake Tahoe, Nevada: Curran Associates Inc., pp. 3111–3119. URL: https://dl.acm.org/doi/10.5555/2999792.2999959.

Miranda, Lester James (2021). *Towards data-centric machine learning: A short review*. URL: https://ljvmiranda921.github.io/notebook/2021/07/30/data-centric-ml.

Mishra, Swaroop, Anjana Arunkumar, Chris Bryan and Chitta Baral (2022). 'A Survey of Parameters Associated with the Quality of Benchmarks in NLP'. In: *ArXiv* abs/2210.07566. URL: https://doi.org/10.48550/arXiv.2210.07566.

Moeller, Sarah (2021). 'Integrating machine learning into language documentation and description'. PhD thesis. University of Colorado. URL: https://www.proquest.com/openview/1de235648809aa25408378d76bd3f99c/1?pq-origsite=gscholar&cbl=18750&diss=y.

Moeller, Sarah and Mans Hulden (Aug. 2018). 'Automatic Glossing in a Low-Resource Setting for Language Documentation'. In: *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 84–93. URL: https://aclanthology.org/W18-4809.

Moeller, Sarah, Ghazaleh Kazeminejad, Andrew Cowell and Mans Hulden (Aug. 2018). 'A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer'. In: *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 12–20. URL: https://aclanthology.org/W18-4802.

Moeller, Sarah, Ling Liu, Changbing Yang, Katharina Kann and Mans Hulden (Nov. 2020). 'IGT2P: From Interlinear Glossed Texts to Paradigms'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5251–5262. DOI: 10.18653/v1/2020.emnlp-main.424. URL: https://aclanthology.org/2020.emnlp-main.424.

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

Nicolai, Garrett, Colin Cherry and Grzegorz Kondrak (June 2015). 'Inflection Generation as Discriminative String Transduction'. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, pp. 922–931. DOI: 10.3115/v1/N15-1093. URL: https://aclanthology.org/N15-1093.

Nicolai, Garrett, Saeed Najafi and Grzegorz Kondrak (Oct. 2018). 'String Transduction with Target Language Models and Insertion Handling'. In: *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Brussels, Belgium: Association for Computational Linguistics, pp. 43–53. DOI: 10.18653/v1/W18-5805. URL: https://aclanthology.org/W18-5805.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira et al. (2016). 'Universal dependencies v1: A multilingual treebank collection'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666. URL: https://aclanthology.org/L16-1262.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers and Daniel Zeman (May 2020). 'Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection'. English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497.

Nzeyimana, Antoine and Andre Niyongabo Rubungo (May 2022). 'KinyaBERT: A Morphology-aware Kinyarwanda Language Model'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 5347–5363. DOI: 10.18653/v1/2022.acl-long.367. URL: https://aclanthology.org/2022.acl-long.367.

O'Keeffe, Anne and Michael McCarthy (2021). *The Routledge handbook of Corpus Linguistics*. 2nd. Routledge London. DOI: https://doi.org/10.4324/9780367076399.

Otter, Daniel W, Julian R Medina and Jugal K Kalita (2020). 'A survey of the usages of deep learning for natural language processing'. In: *IEEE Transactions on Neural Networks and Learning Systems*. URL: https://doi.org/10.1109/TNNLS.2020.2979670.

Otter, Daniel W., Julian R. Medina and Jugal K. Kalita (2021). 'A Survey of the Usages of Deep Learning for Natural Language Processing'. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2, pp. 604–624. DOI: 10.1109/TNNLS.2020.2979670.

Palmer, Alexis (2009). 'Semi-automated annotation and active learning for language documentation'. PhD thesis. University of Texas. URL: https://repositories.lib.utexas.edu/handle/2152/19805.

Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell and Telma Can (2010). 'Computational strategies for reducing annotation effort in language documentation: A case study in creating interlinear texts for Uspanteko'. In: *Linguistic Issues in Language Technology* 3. URL: https://doi.org/10.33011/lilt.v3i.1217.

Park, Hyunji Hayley, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu and Lane Schwartz (2021). 'Morphology Matters: A Multilingual Language Modeling Analysis'. In: *Transactions of the Association for Computational Linguistics* 9. Ed. by Brian Roark and Ani Nenkova, pp. 261–276. DOI: 10.1162/tacl_a_00365. URL: https://aclanthology.org/2021.tacl-1.16.

Pennington, Jeffrey, Richard Socher and Christopher D Manning (2014). 'Glove: Global vectors for word representation'. In: *Proceedings of the 2014 conference on*

empirical methods in natural language processing (EMNLP), pp. 1532–1543. URL: https://aclanthology.org/D14-1162.

Pertsova, Katya (2016). 'Machine Learning of Inflection'. In: *The Oxford Handbook of Inflection*. URL: https://doi.org/10.1093/oxfordhb/9780199591428.013.14.

Piantadosi, Steven T (2014). 'Zipf's word frequency law in natural language: A critical review and future directions'. In: *Psychonomic bulletin & review* 21, pp. 1112–1130. URL: https://doi.org/10.3758/s13423-014-0585-6.

Pimentel, Tiago et al. (Aug. 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*. Online: Association for Computational Linguistics, pp. 229–259. DOI: 10.18653/v1/2021.sigmorphon-1.25. URL: https://aclanthology.org/2021.sigmorphon-1.25.

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova and Anna Korhonen (Sept. 2019). 'Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing'. In: *Computational Linguistics* 45.3, pp. 559–601. ISSN: 0891-2017. DOI: 10.1162/coli_a_00357. eprint: https://direct.mit.edu/coli/article-pdf/45/3/559/1847397/coli_a_00357.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00357.

Raineri, Sophie and Camille Debras (2019). 'Corpora and representativeness: Where to go from now?' In: *CogniTextes. Revue de l'Association française de linguistique cognitive* 19. DOI: https://doi.org/10.4000/cognitextes.1311.

Rastogi, Pushpendre, Ryan Cotterell and Jason Eisner (June 2016). 'Weighting Finite-State Transductions With Neural Context'. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 623–633. DOI: 10.18653/v1/N16-1076. URL: https://aclanthology.org/N16-1076.

Ravfogel, Shauli, Yoav Goldberg and Francis Tyers (Nov. 2018). 'Can LSTM Learn to Capture Agreement? The Case of Basque'. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 98–107. DOI: 10.18653/v1/W18-5412. URL: https://www.aclweb.org/anthology/W18-5412.

Ribeiro, Joana, Shashi Narayan, Shay B. Cohen and Xavier Carreras (Aug. 2018). 'Local String Transduction as Sequence Labeling'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1360–1371. URL: https://aclanthology.org/C18-1115.

Rice, Keren (2006). 'Let the language tell its story? The role of linguistic theory in writing grammars'. In: *Catching Language*. Ed. by Felix K. Ameka, Alan Dench and Nicholas Evans. Berlin, New York: De Gruyter Mouton, pp. 235–268. ISBN: 9783110197693. DOI: doi:10.1515/9783110197693.235. URL: https://doi.org/10.1515/9783110197693.235.

Ruder, Sebastian (2020). *Why You Should Do NLP Beyond English*. http://ruder.io/nlp-beyond-english. URL: https://ruder.%20io/nlp-beyond-english.

Samardžić, Tanja, Robert Schikowski and Sabine Stoll (July 2015). 'Automatic interlinear glossing as two-level sequence classification'. In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Beijing, China: Association for Computational Linguistics, pp. 68–72. DOI: 10.18653/v1/W15-3710. URL: https://aclanthology.org/W15-3710.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh and Lora M Aroyo (2021). '"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI'. In: *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. URL: https://doi.org/10.1145/3411764.3445518.

Seifart, Frank, Nicholas Evans, Harald Hammarström and Stephen C Levinson (2018). 'Language documentation twenty-five years on'. In: *Language* 94.4, e324–e345. URL: https://doi.org/10.1353/lan.2018.0070.

Sennrich, Rico and Barry Haddow (Aug. 2016). 'Linguistic Input Features Improve Neural Machine Translation'. In: *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*. Berlin, Germany: Association for Computational Linguistics, pp. 83–91. DOI: 10.18653/v1/W16-2209. URL: https://aclanthology.org/W16-2209.

Shao, Yan, Christian Hardmeier and Joakim Nivre (July 2018). 'Universal Word Segmentation: Implementation and Interpretation'. In: *Transactions of the Association for Computational Linguistics* 6, pp. 421–435. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00033. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00033/1567640/tacl_a_00033.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00033.

Sharoff, Serge (2017). 'Corpus and systemic functional linguistics'. In: *The Routledge handbook of systemic functional linguistics*. Routledge, pp. 557–570. DOI: https://doi.org/10.4324/9781315413891-46.

Silfverberg, Miikka and Mans Hulden (Oct. 2018). 'An Encoder-Decoder Approach to the Paradigm Cell Filling Problem'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2883–2889. DOI: 10.18653/v1/D18-1315. URL: https://aclanthology.org/D18-1315.

Simons, Gary F and M Paul Lewis (2013). 'The World's languages in crisis'. In: *Responses to language endangerment: In honor of Mickey Noonan. New directions in language documentation and language revitalization* 3, p. 20.

Sinclair, John (1995). 'From theory to practice'. In: *Spoken English on Computer : Transcription, Mark-up and Application*. Ed. by Geoffrey Leech, Greg Myers and Jenny Thomas, pp. 111–122. ISBN: 9781317891055. DOI: https://doi.org/10.4324/9781315843162-16.

Sloetjes, Han and Peter Wittenburg (2008). 'Annotation by category-ELAN and ISO DCR'. In: *6th international Conference on Language Resources and Evaluation (LREC 2008)*. URL: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_60774.

Smit, Peter, Sami Virpioja, Stig-Arne Grönroos and Mikko Kurimo (Apr. 2014). 'Morfessor 2.0: Toolkit for statistical morphological segmentation'. In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 21–24. DOI: 10.3115/v1/E14-2006. URL: https://aclanthology.org/E14-2006.

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen and Trond Trosterud (June 2014). 'Modeling the Noun Morphology of Plains Cree'. In: *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 34–42. DOI: 10.3115/v1/W14-2205. URL: https://aclanthology.org/W14-2205.

Soricut, Radu and Franz Och (June 2015). 'Unsupervised Morphology Induction Using Word Embeddings'. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, pp. 1627–1637. DOI: 10.3115/v1/N15-1186. URL: https://aclanthology.org/N15-1186.

Sutskever, Ilya, Oriol Vinyals and Quoc V Le (2014). 'Sequence to sequence learning with neural networks'. In: *Advances in neural information processing systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K.Q. Weinberger. Vol. 27, pp. 3104–3112. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Sylak-Glassman, John (2016). 'The composition and use of the universal morphological feature schema (unimorph schema)'. In: *Johns Hopkins University*. URL: https://unimorph.github.io/doc/unimorph-schema.pdf.

Sylak-Glassman, John, Christo Kirov, David Yarowsky and Roger Que (July 2015). 'A Language-Independent Feature Schema for Inflectional Morphology'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China, pp. 674–680. DOI: 10.3115/v1/P15-2111.

Tsarfaty, Reut, Dan Bareket, Stav Klein and Amit Seker (2020). 'From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)?' In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7396–7408. DOI: 10.18653/v1/2020.acl-main.660. URL: https://www.aclweb.org/anthology/2020.acl-main.660.

Tyers, Francis, Aziyana Bayyr-ool, Aelita Salchak and Jonathan Washington (2016). 'A finite-state morphological analyser for Tuvan'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2562–2567. URL: https://aclanthology.org/L16-1407.

Vania, Clara, Andreas Grivas and Adam Lopez (Oct. 2018). 'What do character-level models learn about morphology? The case of dependency parsing'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2573–2583. DOI: 10.18653/v1/D18-1278. URL: https://aclanthology.org/D18-1278.

Vania, Clara and Adam Lopez (July 2017). 'From Characters to Words to in Between: Do We Capture Morphology?' In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 2016–2027. DOI: 10.18653/v1/P17-1184. URL: https://www.aclweb.org/anthology/P17-1184.

Vylomova, Ekaterina, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas

Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg and Mans Hulden (July 2020). 'SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection'. In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics, pp. 1–39. DOI: 10.18653/v1/2020.sigmorphon-1.1. URL: https://www.aclweb.org/anthology/2020.sigmorphon-1.1.

Whang, Steven Euijong, Yuji Roh, Hwanjun Song and Jae-Gil Lee (2023). 'Data collection and quality challenges in deep learning: A data-centric ai perspective'. In: *The VLDB Journal*, pp. 1–23. URL: https://doi.org/10.1007/s00778-022-00775-9.

Wiemerslage, Adam, Changbing Yang, Garrett Nicolai, Miikka Silfverberg and Katharina Kann (July 2023). 'An Investigation of Noise in Morphological Inflection'. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 3351–3365. URL: https://aclanthology.org/2023.findings-acl.207.

Williamson, Robert (2020). 'Process and Purpose, Not Thing and Technique: How to Pose Data Science Research Challenges'. In: *Harvard Data Science Review* 3. URL: https://hdsr.mitpress.mit.edu/pub/f2cllynw.

Wu, Shijie, Pamela Shapiro and Ryan Cotterell (Oct. 2018). 'Hard Non-Monotonic Attention for Character-Level Transduction'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4425–4438. DOI: 10.18653/v1/D18-1473. URL: https://aclanthology.org/D18-1473.

Wurm, Stephen A (2001). *Atlas of the World's Languages in Danger of Disappearing*. Unesco.

Xue, Nianwen (Feb. 2003). 'Chinese Word Segmentation as Character Tagging'. In: *International Journal of Computational Linguistics & Chinese Language Processing*,

Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing, pp. 29–48. URL: https://aclanthology.org/O03-4002.

Young, Tom, Devamanyu Hazarika, Soujanya Poria and Erik Cambria (2018). 'Recent trends in deep learning based natural language processing'. In: *IEEE Computational intelligenCe magazine* 13.3, pp. 55–75. DOI: 10.1109/MCI.2018.2840738.

Yu, Yong, Xiaosheng Si, Changhua Hu and Jianxun Zhang (July 2019). 'A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures'. In: *Neural Computation* 31.7, pp. 1235–1270. ISSN: 0899-7667. DOI: 10.1162/neco_a_01199. URL: https://doi.org/10.1162/neco%5C_a%5C_01199.

Zhao, Xingyuan, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig and Lori Levin (Dec. 2020). 'Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations'. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5397–5408. DOI: 10.18653/v1/2020.coling-main.471. URL: https://aclanthology.org/2020.coling-main.471.

Zipf, George Kingsley (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA and London, England: Harvard University Press. ISBN: 9780674434929. DOI: doi:10.4159/harvard.9780674434929. URL: https://doi.org/10.4159/harvard.9780674434929.

Zipf, George Kingsley (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.