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.

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).

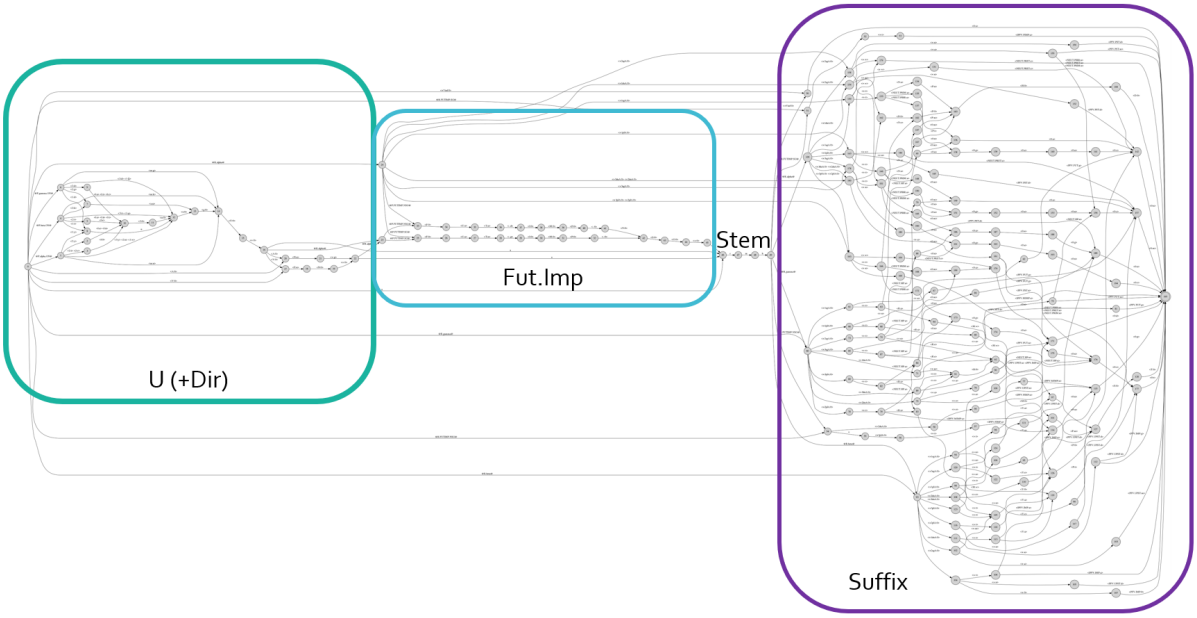

Figure 1: Overall FST architecture for 'Chunking' model. For larger view: 'Chunking'

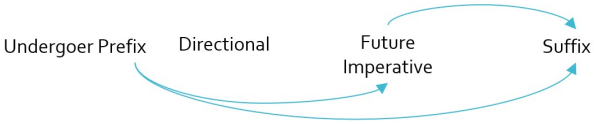

Figure 2: Information flow for the 'Chunking' model.

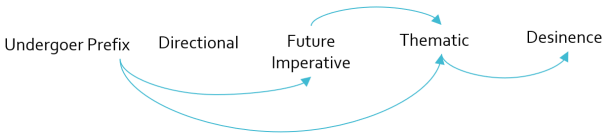

Figure 3: Information flow for decomposition model.

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.

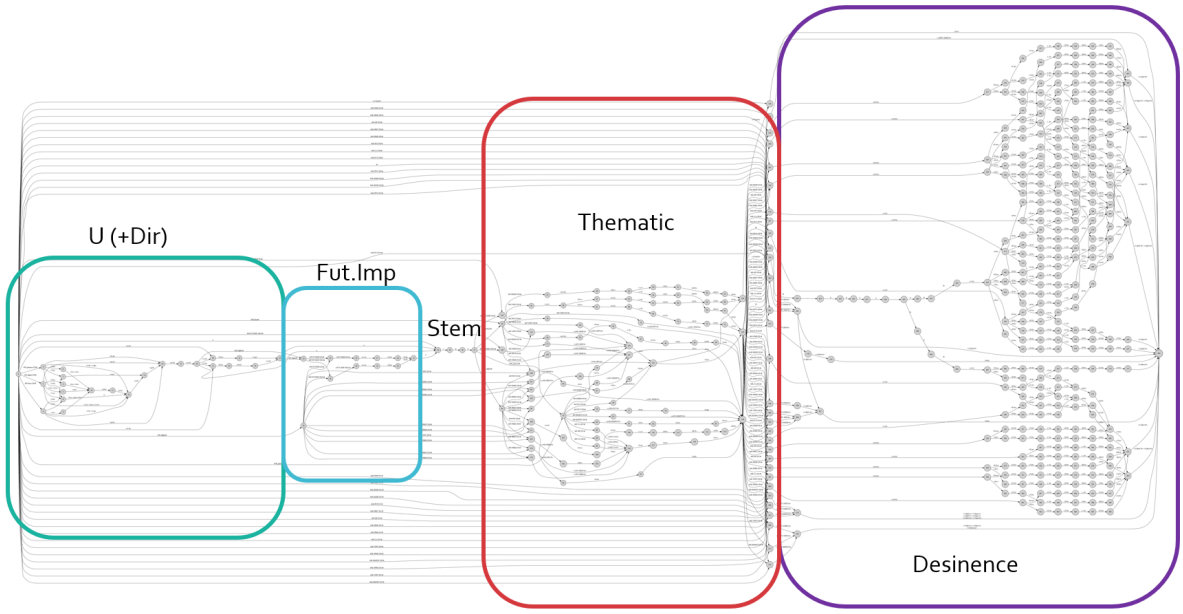

Figure 4: Overall FST architecture for decomposed model. For larger view: Decomposition model.

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.