Unlocking finite-state morphological transducers: Derivational networks for Inuit-Yupik languages

Anonymous ACL submission

1 Introduction

While morphology has received substantial attention in computational linguistics and typology, inflectional resources have long out-classed derivational datasets despite growing interest. UniMorph 4.0 (Batsuren et al., 2022), and Universal Derivations (Kyjánek et al., 2020) contain derivational information for 30 and 21 languages respectively, dwarfed by UniMorph's 169 languages for inflection. Further, the typological diversity of languages covered is still limited and dominated by high-resource (Indo-)European languages, with many of the world's most morphologically rich languages (such as so-called polysynthetic languages) entirely excluded from existing datasets.

While existing derivational datasets are limited in terms of typology and language resource status, there is another, closely related resource available for a much broader array of languages: finite-state morphological transducers (FSMTs). These models encode both lexical and morphological information and exist for a wide range of languages, especially very low-resource, morphologically rich languages. This information is stored in a very different form than existing inflectional and derivational morphological resources, however, and is typically not viewed as a dataset, but as a tool.

In this work, we explore the possibility of using FSMTs to create derivational morphology datasets. We focus on the Universal Derivations (UDer) format. This format is richer than that of UniMorph, capturing not just derivationally-related pairs, but the tree structure of entire derivationally-related families of forms. This makes it particularly suitable for capturing derivational information in highly agglutinative, morphologically-rich languages. In this work, we focus on the Inuit-Yupik language family. These languages are known for having an extremely high degree of synthesis, while being heavily agglutinative, and have frequently been cited as canonical examples of polysynthesis, with a higher type-token ratio than any other language family (Park et al., 2021). Further, several languages in the family (`kal`, `ess`, `iku`, `esu`) have FSMTs publicly available. We produce Universal Derivations-style datasets for Greenlandic (`kal`; ∼44,000 speakers) and Saint Lawrence Island (SLI) Yupik (`ess`; ∼500 speakers), using publicly available FSMTs and small text corpora. We make our code and derivational networks in Universal Dependencies format available online.[1]

2 Method

Most FSMTs are primarily designed for morphological analysis; as such, they may generate forms which, while seemingly valid, do not occur (e.g. paradigm gaps). To avoid including such items in our derivational networks, we use existing text corpora for the two languages and use the FSTMs to *analyse* these corpus–thereby restricting us to attested surface forms. We use the digital corpus of SLI Yupik[2], consisting of ∼300,000 unannotated tokens and ∼1,000 manually annotated tokens, and the monolingual Greenlandic corpus collected by Jones (2022), comprising 1.98 million tokens. We use (Chen and Schwartz, 2018)'s FSMT for SLI Yupik and the Apertium morphological analyser for Greenlandic to provide morphological analyses for the corpora (Forcada and Tyers, 2016).

As described in Figure 1, our method works by first analyzing words in the corpus, then repeatedly modifying the analysis and generating forms matching the modified analysis. Because these languages are exclusively suffixing, we can therefore recursively strip off derivational morphemes without worrying about ordering.

[1] URL added upon acceptance.

[2] https://github.com/SaintLawrenceIslandYupik/digital_corpus

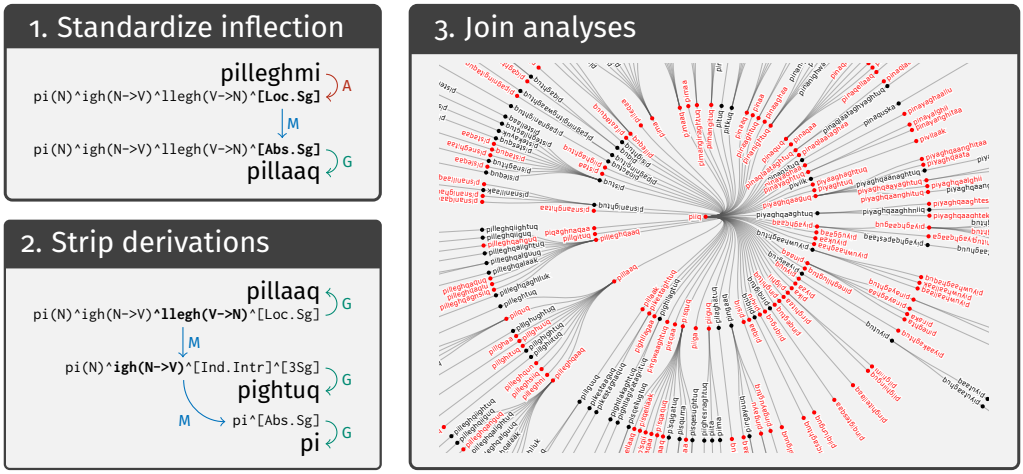

Figure 1: Our method for producing derivational networks from FSMTs. Words in a corpus are first analyzed (A) using the FSMT. We then modify (M) the analysis to have standard inflectional features, and then generate (G) the standardized form with the FSMT. Next, we recursively modify to strip derivations and generate intermediate forms, producing a chain of derivationally related words. We join chains of derivationally related words to form a network. Red lexemes are attested in the corpus, while black forms are inferred from attested derived forms.

3 Results

Our derivational networks cover 53,245 lexemes for SLI Yupik and 127,663 lexemes for Greenlandic, on par or surpassing highly-resourced European languages such as Dutch, French, Italian, and English.Further, these lexemes are spread across 6,344 (SLI Yupik) and 11,088 (Greenlandic) distinct derivational families. In contrast to less rich languages, a *majority* of these families are non-trivial (containing at least two lexemes): 4,256 and 6,021; respectively. Further, in both languages almost 1 in 10 derivational families contained 20 or more lexemes (599 `ess`; 1,015 `kal`). The largest derivational families in each language contain many hundreds of lexemes: 359 for the neutral root *piiq* in SLI Yupik, and 1,584 for Greenlandic, far surpassing any single lexeme in existing UDer languages. Finally, we note an impressive range of unique derivational relations/morphemes covered: 397 in SLI Yupik and 327 in Greenlandic.

While this data cannot be considered gold-standard, existing FSMTs and small corpora can yield large, empirically-grounded derivational networks for extremely low-resource morphologically rich languages. These networks could serve to speed up native speaker annotation, or as silver-standard data in certain types of analysis. These findings corroborate the noted derivational richness of Inuit-Yupik languages. Future work could focus on improving these networks, extending to other languages, building tools for human annotators, or refining these techniques for language with ambiguous morpheme sequencing or parts of speech.

References

Khuyagbaatar Batsuren et al. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mikel L. Forcada and Francis M. Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Alex Jones. 2022. Finetuning a Kalaallisut-English machine translation system using web-crawled data. *Preprint*, arXiv:2206.02230.

Lukáš Kyjánek et al. 2020. Universal Derivations 1.0, A growing collection of harmonised word-formation resources. *Prague Bulletin of Mathematical Linguistics*, 115(1):5–30.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.