# Data-Driven Morphological Analysis for Uralic Languages

Miikka Silfverberg
University of Helsinki
Department of Digital Humanities
00014-FI Helsinki
miikka.silfverberg@helsinki.fi

Francis M. Tyers
Indiana University
Department of Linguistics
Bloomington, IN 47408
ftyers@indiana.edu

**Abstract**

This paper describes an initial set of experiments in data-driven morphological analysis of Uralic languages. The paper differs from previous work in that our work covers both lemmatization and generating ambiguous analyses. While hand-crafted finite-state transducers represent the state of the art in morphological analysis for most Uralic languages, we believe that there is a place for data-driven approaches, especially with respect to making up for lack of completeness in the ɯlexicon. We present results for nine Uralic languages that show that, at least for basic nominal morphology for six out of the nine languages, data-driven methods can achieve an F-score of over 90%, providing results that approach those of finite-state techniques. We also compare our system to an earlier approach to Finnish data-driven morphological analysis (Silfverberg and Hulden, 2018) and show that our system outperforms this baseline.

**Abstract**

Tämä artikkeli esittelee kokeita uralilaisten kielten morfologisessa analyysissä koneoppimismenetelmin. Artikkeli eroaa aiemmista lähestymistavoista, koska se tuottaa lemmoja morfologisten analyysien osana ja pystyy tuottamaan useampia analyysejä monitulkintaisille sanoille. Vaikka sääntöpojaiset käsin tehdyt analysaattorit vielä selkeästi päihittävät koneoppimismenetelmin rakennetut analysaattorit, uskomme että koneoppimismenetelmillä on sija morfologisen analyysin alalla varsinkin perinteisten analysaatorien sovellusalan kasvattamisessa. Tässä artikkelissa esittelemme koetuloksia yhdeksälle uralilaiselle kielelle. Osoitamme että on mahdollista oppia analysoimaan substantiivien perusmorfologiaa 90% F1-score tasolla, mikä lähestyy olemassa olevia sääntöpohjaisten järjestelmien tasoa. Vertaamme myös järjestelmäämme aiemmin esiteltyyn koneoppimismenetelmin rakennettuun morfologiseen jäsentimeen (Silfverberg and Hulden, 2018) ja osoitamme, että meidän järjestelmämme on parempi.





*Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages*, pages 1–14,
Tartu, Estonia, January 7 - January 8, 2019.

# 1 Introduction

Morphological analysis is the task of producing, for a given surface form, a list of all and only the valid analyses in the language. For example, given the surface form *voisi* in Finnish, a morphological analyser must produce not only the most frequent analysis `voida+VERB|Mood=Cond|Number=Sg|Person=3` 'can they?' but also the less frequent `voida+VERB|ConNeg=Yes` 'can-NEG', or theoretical/rare ones `voi+NOUN|Number=Sg|Case=Nom|Possessor=Sg2` 'your butter'.

Morphological analysis is a cornerstone of language technology for Uralic and other morphologically complex languages, where type-to-token ratio becomes prohibitive for purely word based methods. Rule-based morphological analyzers (Beesley and Karttunen, 2003) represent the current state-of-the-art for this task. The analyses returned by such systems are typically very accurate, however, rule-based systems suffer from low coverage since novel lexical items often need to be manually added to the system.[1]

We explore the task of data-driven morphological analysis, that is, learning a model for analyzing previously unseen word forms based on a morphologically analyzed text corpus. This can help with the coverage problem encountered with rule-based analyzers. Morphological guessers based on existing rule-based analyzers represent a classical approach to extending the coverage of a rule-based analyzer. These are constructed by transforming an original analyzer typically using weighted finite-state methods (Lindén, 2009). In practice, this limits the range of data-driven models that can be applied. For example, models which do not incorporate a Markov assumption (such as RNNs) can be difficult to apply due to the inherent finite-state nature of rule-based analyzers.

Our system[2] is a neural encoder-decoder which is learned directly from morphologically analyzed text corpora. It is inspired by previous approaches to morphological analysis by Moeller et al. (2018) and Silfverberg and Hulden (2018). In contrast to these existing neural morphological analyzers, our system produces full morphological analyses: it provides both morphological tags and lemmas as output and it can return multiple alternative analyses for one input word form using beam search.

We present experiments on morphological analysis of nouns for nine Uralic languages: Estonian, Finnish, Komi-Zyrian, Moksha, Hill Mari, Meadow Mari, Erzya, North Sámi and Udmurt. We show that our system achieves roughly 90% F1-score for most of the tested languages. Additionally, we compare our system to the Finnish data-driven morphological analyzer presented by Silfverberg and Hulden (2018). As seen in Section 5, our system clearly outperforms the earlier approach.

# 2 Related Work

There is a strong tradition of work on rule-based morphological analysis for Uralic languages. Recent examples include Pirinen et al. (2017), Trosterud et al. (2017) and Antonsen et al. (2016), although work in the area has been going on for many years (cf. Koskenniemi (1983)). There is also a growing body of work on data-driven morphological tagging for Uralic languages, especially Finnish. Here, a system is trained to find a single contextually appropriate analysis for each token in a text. Examples of

[1] Although novel lexical items can cause problems for data-driven systems as well, most data-driven systems are still able to analyze any word form in principle.

[2] Code available at `https://github.com/mpsilfve/morphnet`.

work exploring morphological tagging for Finnish include Kanerva et al. (2018) and Silfverberg et al. (2015). However, work on full data-driven morphological analysis, where the task is to return all and only the valid analyses for each token irrespective of sentence context, is almost non-existent for Uralic languages. The only system known to the authors is the recent neural analyzer for Finnish presented by Silfverberg and Hulden (2018). The system first encodes an input word form into a vector representation using an LSTM encoder. It then applies one binary logistic classifier conditioned on this vector representation for each morphological tag (for example `NOUN|Number=Sg|Case=Nom`). The classifier is used to determine if the tag is a valid analysis for the given input word form. Similarly to Silfverberg and Hulden (2018), our system is also a neural morphological analyzer but unlike Silfverberg and Hulden (2018) we incorporate lemmatization. Moreover, the design of our system considerably differs from their system as explained below in Section 3.

The lack of work on morphological analysis for Uralic languages is unsurprising because the field of data-driven morphological analysis in general remains underexplored at the present time. Classically, morphological analyzers have been extended using morphological guessers (Lindén, 2009), however, the premise for such work is quite different—An existing analyzer is modified to analyze unknown word forms based on orthographically similar known word forms. In contrast, we explore a setting, where the starting point is a morphologically analyzed corpus and the aim is to learn a model for analyzing unseen text.

Outside of the domain of Uralic languages, Nicolai and Kondrak (2017) frame morphological analysis as a discriminative string transduction task. They present experiments on Dutch, English, German and Spanish. In contrast to Nicolai and Kondrak (2017), Moeller et al. (2018) use a neural encoder-decoder system for morphological analysis of Arapaho verbs. Their system returns both lemmas and morphological tags but it cannot handle ambiguous analyses in general.[3] Our work is inspired by the neural encoder-decoder approach presented by Moeller et al. (2018) but we do handle unrestricted ambiguity.

In contrast to data-driven morphological analysis, data-driven morphological generation has received a great deal of attention lately due to several shared tasks organized by CoNLL and SIGMORPHON (Cotterell et al., 2016, 2017, 2018). The most successful approaches (Kann and Schütze, 2016; Bergmanis et al., 2017; Makarov et al., 2017; Makarov and Clematide, 2018) to the generation task involve different flavors of the neural encoder-decoder model. Therefore, we opted for applying it in our morphological analyzer.

# 3 Model

This section presents the encoder-decoder model used in the experiments.

## 3.1 An Encoder-Decoder Model for Morphological Analysis

Following Moeller et al. (2018), we formulate morphological analysis as a character-level string transduction task and use an LSTM (Hochreiter and Schmidhuber, 1997) encoder-decoder model with attention (Bahdanau et al., 2014) for performing the string transduction. To this end, we train our model to translate input word forms

[3]The system can handle ambiguity in limited cases by using underspecified tags. For example an ambiguity between singular and dual number could be expressed using a tag `[SG/DPL]`.

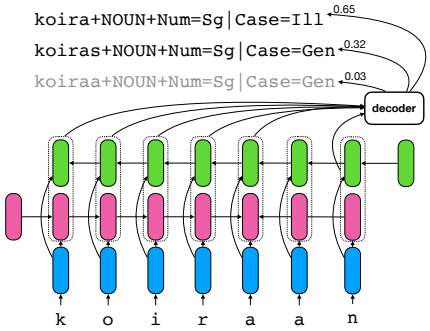


**Figure 1:** We use a bidirectional LSTM encoder for encoding an input word form into forward and backward states (pink and green bars) one character at a time. We then use an attentional LSTM decoder for generating output analyses one symbol at a time. We return the least number of most probable analyses whose combined mass is greater than a threshold $p$. In this example, for $p = 0.9$, the analyzer would return `koira+NOUN+Num=Sg|Case=Ill` and `koiras+NOUN+Num=Sg|Case=Gen` whose combined probability mass is $0.97$, given the input form `koiraan`.

like *koiraan* (singular illative for *koira* 'dog' or singular genitive for *koiras* 'male' in Finnish) into a set of output analyses:

```
koira+NOUN+Number=Singular|Case=Ill
koiras+NOUN+Number=Singular|Case=Gen
```

Each analysis consists of a lemma (`koira` 'dog'), a part-of-speech (POS) tag (`NOUN`) and a morphosyntactic description (MSD) (`Number=Singular|Case=Gen`). The procedure is illustrated in Figure 1.

Above, we presented an example from Finnish, *voisi*, which can be both an inflected form of a noun and an inflected form of a verb. This shows that a word form may have multiple valid morphological analyses with different lemmas, POS tags and MSDs. Therefore, our model needs to be able to generate multiple output analyses given an input word form. We accomplish this by extracting several output candidates from the model using beam search and selecting the most probable candidates as model outputs. The number of outputs is controlled by a probability threshold hyperparameter $p$. We extract the least number of top scoring candidates whose combined probability mass is greater than $p$. Additionally, we restrict the maximal number of output candidates using a single hyperparameter $N$. The hyperparamaters $p$ and $N$ are tuned on the development data.

## 3.2 Implementation Details

We implement our LSTM encoder-decoder model using the OpenNMT neural machine translation toolkit (Klein et al., 2017). We use 500-dimensional character and tag embeddings for input and output characters as well as POS and MSD tags. These are processed by a 2-layer bidirectional LSTM encoder with hidden state size 500. Encoder representations are fed into a 2-layer LSTM decoder with hidden state size 500. During inference, we use beam search with beam width 10.

When training, we use a batch size of 64 and train for 10,000 steps where one step corresponds to updating on a single mini-batch. Model parameters are optimized using the Adam optimization algorithm (Kingma and Ba, 2014).

# 4 Data

We use two datasets in the experiments. The first dataset is created by using the morphological transducers from Giellatekno to analyze wordforms in a frequency list from Uralic Wikipedias. The second one is created using data from the Turku Dependency Treebank. This dataset was originally presented by Silfverberg and Hulden (2018). We explicitly do not use any data from the Unimorph project.

## 4.1 Uralic Wikipedia Data

We applied the models to nine Uralic languages: Erzya (`myv`), Estonian (`est`), Finnish (`fin`), Komi-Zyrian (`kpv`), Hill Mari (`mhr`), Meadow Mari (`mrj`), Moksha (`mdf`), North Sámi (`sme`) and Udmurt (`udm`). These languages were chosen as they had both a moderately-sized free and open text corpus (Wikipedia) and an existing free/open-source morphological analyser from the Giellatekno infrastructure (Moshagen et al., 2014). Hungarian (`hun`) was omitted as there was no functional analyser in the Giellatekno infrastructure, while the remainder of the Sámi languages (i.e. South (`sma`), Lule (`smj`), Inari (`smn`), ...) and Kven (`fkv`) were left out as they have as yet no Wikipedia. The remainder of the Uralic languages have neither wide-coverage analyser nor Wikipedia.

The data used in the experiments consisted of tab separated files with five columns: language code, surface form, lemma, part-of-speech and list of morphological tags expressed as Feature=Value pairs (see Figure 2). Both the parts of speech and the morphological tags broadly follow the conventions of the Universal Dependencies project (Nivre et al., 2016), with one exception: The tags are given in the same order they appear in the original morphological analyses (largely morpheme order) as opposed to in alphabetical order by feature name.

Each file was generated as follows: First we downloaded the relevant Wikipedia dump[4] and extracted the text using `WikiExtractor`.[5] This gave us a plain-text corpus of the language in question. We then used the morphological transducers from Giellatekno (Moshagen et al., 2014) to both tokenize and analyze the text. This was then made into a frequency list using standard Unix utilities. We then extracted only the forms with noun analyses and removed all non-noun analyses, along with noun analyses that included numerals, abbreviations, acronyms, spelling errors or dialectal forms. All derived and compound analyses were also removed, in addition to analyses that included clitics (e.g. Finnish *-kään, -kaan*). The exclusion of these phenomena makes the task less applicable to a real-world setting, but at the same time makes it tractable for initial experiments such as the ones presented in this paper.

After creating the frequency list, we converted the format of the analyses by means of a simple lookup table (e.g. `+Gen` → `Case=Gen`). An example from the training data of North Sámi can be found in Figure 2 and details about the size of the training data for each of the languages can be found in Table 1.

All data sets were randomly split into 80% training data, 10% development data and 10% test data. The splits are disjoint in the sense that the training and development set never include word forms seen in the test set. They may, however, include other inflected forms of lemmas that do occur in the test set.

[4] Available from `https://dumps.wikimedia.org`.

[5] `https://github.com/apertium/WikiExtractor`

```
sme    čuđiid    čuhti    NOUN    Number=Plur|Case=Acc
sme    čuđiid    čuhti    NOUN    Number=Plur|Case=Gen
sme    čuđiid    čuohti   NOUN    Number=Plur|Case=Acc
sme    čuđiid    čuohti   NOUN    Number=Plur|Case=Gen
sme    čuđiid    čuđđi    NOUN    Number=Plur|Case=Acc
sme    čuđiid    čuđđi    NOUN    Number=Plur|Case=Gen
```

**Figure 2:** Example from the North Sámi (`sme`) training data for the forms of the word *čuđiid*, which could be a form of *čuhti* 'Chud', *čuohti* 'hundred' or *čuđđi* 'enemy'.

| Language | Code | Data | | | Ambig. | Lemmas | Tags | MSDs | Th. forms |
|---|---|---|---|---|---|---|---|---|---|
| | | `train` | `dev` | `test` | | | | | |
| Estonian | `est` | 87930 | 10991 | 10991 | 1.11 | 17814 | 16 | 26 | 32 |
| Finnish | `fin` | 153603 | 19200 | 19200 | 1.17 | 44644 | 21 | 142 | 180 |
| Komi-Zyrian | `kpv` | 9413 | 1176 | 1176 | 1.44 | 3602 | 25 | 113 | 312 |
| Moksha | `mdf` | 3456 | 431 | 431 | 1.16 | 1479 | 23 | 66 | 156 |
| Hill Mari | `mhr` | 7789 | 973 | 973 | 1.11 | 4650 | 17 | 20 | 120 |
| Meadow Mari | `mrj` | 6885 | 860 | 860 | 1.11 | 2923 | 13 | 55 | 108 |
| Erzya | `myv` | 11384 | 1423 | 1423 | 1.50 | 5204 | 23 | 90 | 288 |
| North Sámi | `sme` | 9328 | 1166 | 1166 | 1.90 | 6032 | 16 | 54 | 126 |
| Udmurt | `udm` | 6344 | 792 | 792 | 1.23 | 3115 | 23 | 100 | 150 |

**Table 1:** Quantitative description of the Uralic Wikipedia datasets. For each language, we give its ISO 639-3 code (**Code**), the number of unique `train`, `dev` and `test` word forms, as well as, the average number of analyses per word form (**Ambig.**), the number of unique lemmas (**Lemmas**), the number of unique tags such as `NOUN` and `Number=Sing` (**Tags**) and the number of unique morphosyntactic descriptions such as `NOUN|Number=Sing|Case=Nom` (**MSDs**) in the dataset. In addition we provide an approximate number of possible theoretical forms in the noun paradigm for each language (**Th. forms**). Note that both the **Ambig.** and **Th. forms** columns give the theoretical maximum only for the morphological features of number, case and possession and does not include forms generated by productive derivation or addition of clitics.

| Language | Code | Data | | | Ambig. | Lemmas | Tags | MSDs | Th. forms |
|---|---|---|---|---|---|---|---|---|---|
| | | train | dev | test | | | | | |
| Finnish UD | — | 162827 | 18311 | 21070 | 1.80 | — | 137 | 2452 | — |

**Table 2:** Quantitative description of the Finnish treebank dataset. We give the number of unique `train`, `dev` and `test` word forms, as well as, the average number of analyses per word form (**Ambig.**), the number of unique lemmas (**Lemmas**), the number of unique tags such as `NOUN` and `Number=Sing` (**Tags**) and the number of unique morphosyntactic descriptions such as `NOUN|Number=Sing|Case=Nom` (**MSDs**).

### 4.2 Finnish Treebank Data

Our second dataset was presented by Silfverberg and Hulden (2018). It is the Finnish part of the Universal Dependencies treebank v1 (Pyysalo et al., 2015) which has been analyzed using the OMorFi morphological analyzer (Pirinen et al., 2017). We used the splits into training, development and test sets provided by Silfverberg and Hulden (2018).

In contrast to the Uralic Wikipedia datasets, which is a type-level resource consisting of analyses for unique word forms, the Finnish treebank data is a token-level resource consisting of morphologically analyzed running text. Therefore, the same word form can occur multiple times in the dataset. This means that the training, development and test sets are not disjoint which makes the task somewhat easier. However, the dataset contains word forms from all Finnish word classes. It also contains derivations and clitics. This, in turn, makes it more versatile than the Uralic Wikipedia data. The dataset is described in Table 2.

## 5 Experiments and Results

We present results for two experiments. In the first experiment, we train analyzers for the Uralic Wikipedia data presented in Section 4. In the second experiment, we train an analyzer on the Finnish Treebank data used by Silfverberg and Hulden (2018) and compare our system to theirs.

Because an input word form can have several alternative analyses, we present results for precision, recall and F1-score on analyses. These are defined with regard to the quantities *true positives* ($tp$) which is the number of gold standard analyses that our system recovered, *false positives* ($fp$) which is the number of incorrect analyses that our system produced and *false negatives* ($fn$) which is the number of gold standard analyses which our system was unable to recover. Definitions for recall, precision and F1-score are given below:

$$\text{Recall} = \frac{tp}{tp+fn}, \text{Precision} = \frac{tp}{tp+fp} \text{ and F1-score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

### 5.1 Experiment on Uralic Wikipedia Data

We present three different evaluations of the results. Table 3 shows results on complete analyses including the lemma, POS tag and MSD. The results are above 90% F1-score for most languages. The exception to this are Finnish and Northern Sámi where results fall below 90%. Recall is higher than precision for most languages with the exception of Northern Sámi and Komi-Zyrian. Our model achieves the best F1-score for Udmurt (95.09%)

| Language | Recall | Precision | F1-Score |
|---|---|---|---|
| est | 92.97 | 89.26 | 91.08 |
| fin | 89.30 | 87.37 | 88.32 |
| kpv | 88.13 | 92.14 | 90.09 |
| mdf | 93.91 | 90.78 | 92.32 |
| mhr | 89.98 | 89.48 | 89.73 |
| mrj | 93.95 | 91.65 | 92.78 |
| myv | 91.19 | 89.79 | 90.48 |
| sme | 85.95 | 87.12 | 86.53 |
| udm | 96.41 | 93.81 | 95.09 |

**Table 3**: Results for full analyses (lemma + POS + MSD) on the Uralic Wikipedia data.

| Language | Recall | Precision | F1-Score |
|---|---|---|---|
| est | 95.69 | 91.06 | 93.32 |
| fin | 93.41 | 87.72 | 90.48 |
| kpv | 96.51 | 92.81 | 94.62 |
| mdf | 95.15 | 91.96 | 93.53 |
| mhr | 90.69 | 89.84 | 90.26 |
| mrj | 93.71 | 91.25 | 92.46 |
| myv | 95.40 | 90.91 | 93.10 |
| sme | 91.47 | 87.45 | 89.42 |
| udm | 97.26 | 94.32 | 95.77 |

**Table 4**: Results for lemmas on the Uralic Wikipedia data.

| Language | Recall | Precision | F1-Score |
|---|---|---|---|
| est | 95.21 | 94.04 | 94.62 |
| fin | 93.62 | 95.11 | 94.36 |
| kpv | 90.43 | 95.10 | 92.70 |
| mdf | 96.95 | 95.78 | 96.36 |
| mhr | 96.28 | 95.32 | 95.80 |
| mrj | 97.65 | 96.72 | 97.18 |
| myv | 94.62 | 94.39 | 94.50 |
| sme | 92.12 | 94.30 | 93.20 |
| udm | 97.84 | 96.25 | 97.03 |

**Table 5**: Results for tags (POS + MSD) on the Uralic Wikipedia data.

| Error category | Count | % Total |
|---|---|---|
| 1. Loan words in *-ие*, *-ье* or *-ья* | 18 | 18.6 |
| 2. Other Russian loan word mistake | 2 | 2.1 |
| 3. Plural morpheme part of stem | 6 | 6.2 |
| 4. Loan words ending in *ь* | 10 | 9.7 |
| 5. Overenthusiastic lemmatization | 2 | 2.1 |
| 6. Underenthusiastic lemmatization | 3 | 3.1 |
| 7. Impossible lemma | 4 | 4.1 |
| 8. Words containing hyphen | 6 | 6.2 |
| 9. Other | 46 | 47.4 |
| **Total**: | 97 | 100.0 |

**Table 6**: Qualitative evaluation of the errors in the output of the system for Udmurt. The majority of errors can be classified with loan words from Russian making a good proportion.

Table 4 shows results for plain lemmas without POS or MSD. Here all languages except Northern Sámi receive F1-score over 90% and, as in the case of full analyses, recall is again higher than precision for all languages. For lemmas, the best F1-score is again attained on Udmurt (95.77%)

The final evaluation is shown in Table 5. This table shows results for POS and MSD tag. Overall results here are higher than for the lemma or full analysis: in excess of 92% for all languages. Similarly as in the case of full analyses and lemmas, our model again delivers the best F1-score for Udmurt (97.03%).

For the Udmurt data, given that only 97 analyses were incorrect we were able to do a partial qualitative evaluation shown in Table 6. We looked at all the analyses and categorised them into nine error classes: (1) Russian loan words ending in *-ие*, *-ье* or *-ья* that do not receive the right lemma; (2) Other mistakes in loan words from Russian; (3) Plural morpheme is considered part of the stem; (4) Words ending in soft sign *-ь* that were mislemmatized; (5) Overenthusiastic lemmatization — i.e. the system produced a lemma that did not exist in the data; (6) Under enthusiastic lemmatization — i.e. a lemma in the data was not produced by the system; (7) Impossible lemma — i.e. the singular nominative did not have the same form as the lemma; (8) Words with hyphen in; (9) Other.

A typical error of the first type can be found in the lemmatization of the word *путешествие* 'travel', the lemma given by the network was **путешестви*,[6] similarly **междометия* was given for *междометие* 'interjection'. The second error class included errors like the lemma **республик* for the form *республиказы* 'to/in our republic'. The system also sometimes generated lemmas in the plural form (third error type), for example *бурдъёсаз* 'on/to its wings' generated two correct analyses with the lemma *бурд* 'wing' and one incorrect with the lemma *бурдъёс* 'wings'. For errors of the fourth type we can consider the form *пристане* 'wharf-ILL' which has the lemma *пристань* 'wharf' (as in Russian), but for which the system produced both **пристан* and **пристане*, neither of which exist as lemmas in Udmurt or Russian.

For the fifth type we have *спортэ* giving the lemma **спор*[7] instead of *спорт* 'sport'

[6]According to some Udmurt authors this is the preferred nominative singular form, but we count it as an error as the analyser we based the gold standard on uses *путешествие* as the lemma.

[7]Note that this could potentially be a loan of *спор* 'dispute, argument' from Russian, but as it was not in the gold standard counted it as an error.

| System | Recall | Precision | F1-Score |
|---|---|---|---|
| Our System | 93.45 | 96.47 | 94.94 |
| Silfverberg and Hulden (2018) | 89.66 | 94.03 | 91.79 |

**Table 7**: Comparison between our system and Silfverberg and Hulden (2018). We present results for tags only (POS + MSD) since the system by Silfverberg and Hulden (2018) does not lemmatize.

and sixth type *берлань* we get the noun lemma *берлань* instead of *бер* 'back-APPROX'. Note that there is a much more frequent reading of *берлань* as an adverb 'ago, back' (rus. *назад*), but as this was not a nominal reading it was excluded from the experiments.

For the seventh type consider the word *пияш* 'boy, lad' which generated nominative singular analyses with the lemmas *пи* 'son' and **пиеш*.

The system was also confused by compound words written with a hyphen (error type 8). Three out of seven of these had various different kinds of errors, for example losing part of the compound *ваньмыз-ӧвӧлэз* → *ваньмыз*, making compound-internal vowel changes *тодон-эскеронъя* → **тодон-ӧскерон* or considering an affix part of the lemma *музей-коркан* 'village-house museum' → *музей-коркан*.

While the 'Other' class makes up almost half of the data, we can see that over half of the errors should in principle be able to be solved with simply adding more data. That is, the model has not received enough information about how Russian loan words, or words with hyphens work as they compose a small fraction of the data.

### 5.2 Experiment on Finnish Treebank Data

In our second experiment, we compare our system against the neural morphological analyzer proposed by Silfverberg and Hulden (2018). We trained a morphological analyzer on the Finnish treebank training data used by Silfverberg and Hulden (2018) and report results on their test data. Similarly to Silfverberg and Hulden (2018), we also return the set of analyses seen in the training set for those test word forms which were seen in the training data.

Table 7 shows results on the Finnish treebank dataset. We only report results for precision, recall and F1-score with regard to tags (POS + MSD) because the system by Silfverberg and Hulden (2018) is not capable of lemmatization. As Table 7 shows, our system clearly outperforms the system proposed by Silfverberg and Hulden (2018) with regard to F1-Score on tags. Results on the Finnish treebank data are also far better than results on the Finnish Wikipedia data.

## 6 Discussion and Conclusions

On the Uralic Wikipedia data, F1-score for full analyses ranges from 86% for Northern Sámi to 95% for Udmurt with most languages receiving an F1-score around 90%. The weaker performance on Northern Sámi is understandable since the language is known to have an intricate system of morphophonological alternations (see for example the description in Sammallahti (1998)).

Our system clearly outperforms the system by Silfverberg and Hulden (2018) on the Finnish Wikipedia data. In contrast to the Uralic Wikipedia data, the Finnish Treebank dataset, represents continuous text with word forms belonging to a mix of word classes. It also covers clitics and derivations which are missing from the Uralic Wikipedia dataset.[8] Therefore, this experiment indicates that our system is also applicable to analysis of running text for Finnish.

The overall better performance on the Finnish Treebank dataset is explained by the fact that it is a token-level resource where frequent words, which are easy to analyze, can substantially improve performance.

In contrast to what Silfverberg and Hulden (2018) found, our results on the Finnish Wikipedia data indicate that recall is higher than precision for most languages. However, on the Finnish treebank data, we also get higher precision than recall although our system delivers more balanced recall and precision than the system proposed by Silfverberg and Hulden (2018). It is not immediately clear, why it is advantageous to prefer precision over recall but this may be related to the large number of possible POS + MSD combinations in the Finnish Treebank dataset. Many of these could potentially be applicable judging purely on the basis of the orthographical form of a particular word form but only a small number of the combinations will actually result in a valid analysis. Therefore, it may be advantageous to return a more restricted set of highly likely analyses.

As there are other treebanks for Uralic languages, i.e. Hungarian (Vincze et al., 2010), Estonian (Muischnek et al., 2016), North Sámi (Sheyanova and Tyers, 2017) and Erzya (Rueter and Tyers, 2018), we would like to run the equivalent experiments as on the Finnish treebank.

As explained in Section 3, we return analyses based on probability mass. It could be better to predict how many forms are going to be included based on the input word form. For example, if the input word form is markedly different than most forms seen in the training data, the model may assign lower confidence to output analyses. Applying a probability mass threshold in this case may result in a very large number of outputs.

Large training sets are available for only a few Uralic languages, Therefore, we should explore using a hard attention model similar to Makarov and Clematide (2018) in our encoder-decoder. The results from CoNLL SIGMORPHON shared tasks (Cotterell et al., 2018) show that a hard attention model can be a far stronger learner in a low-resource setting.

We presented a data driven morphological analyzer and evaluated its performance on morphological analysis of nouns for nine Uralic languages. Moreover, we evaluated the performance on Finnish running text. Our system delivers encouraging results. F1-score for analysis of nouns is around 90% for most of our languages. In addition, our system substantially improves upon the baseline presented by Silfverberg and Hulden (2018). In future work, we need to explore hard attention models for morphological analysis since these deliver strong performance in low-resource settings which are typical for Uralic languages. Moreover, we need to explore more principled ways to handle ambiguous analyses.

[8]Recall that clitics and derivations are missing as they were removed during processing of the Wikipedia data (§4) to make the data easier to process and more comparable cross-linguistically, as clitics are treated differently in the different analysers.

## Acknowledgements

We wish to thank the anonymous reviewers for their insightful comments. The first author has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

## References

Lene Antonsen, Trond Trosterud, Marja-Liisa Olthuis, and Erika Sarivaara. 2016. Modelling the Inari Saami morphophonology as a finite state transducer. In *The Second International Workshop on Computational Linguistics for Uralic Languages, Szeged, January 2016*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite state morphology*. CSLI publications.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Association for Computational Linguistics, pages 31–39. https://doi.org/10.18653/v1/K17-2002.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. *arXiv preprint arXiv:1810.07125* .

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, et al. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. *arXiv preprint arXiv:1706.09031* .

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Brussels, Belgium, pages 133–142. http://www.aclweb.org/anthology/K18-2013.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 62–70. https://doi.org/10.18653/v1/W16-2010.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints* .

Kimmo Koskenniemi. 1983. Two-level model for morphological analysis. In *IJCAI*. volume 83, pages 683–685.

Krister Lindén. 2009. Guessers for finite-state transducer lexicons. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*. Springer-Verlag, Berlin, Heidelberg, CICLing '09, pages 158–169.

Peter Makarov and Simon Clematide. 2018. UZH at CoNLL-SIGMORPHON 2018 shared task on universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection* pages 69–75.

Peter Makarov, Tatiana Ruzsics, and Simon Clematide. 2017. Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Association for Computational Linguistics, pages 49–57. https://doi.org/10.18653/v1/K17-2004.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 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*. pages 12–20.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. In *Proceedings of the 1st Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL-2014)*. pages 71–77.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies. In *Proceedings of LREC 2016*.

Garrett Nicolai and Grzegorz Kondrak. 2017. Morphological analysis without expert annotation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. volume 2, pages 211–216.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Chris Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference (LREC'16)*.

Tommi A Pirinen, Inari Listenmaa, Ryan Johnson, Francis M. Tyers, and Juha Kuokkala. 2017. Open morphology of Finnish. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11372/LRT-1992.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *NODALIDA*.

Jack Rueter and Francis M. Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. In *Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages*. pages 108–120.

Pekka Sammallahti. 1998. Saamic. In Daniel Abondolo, editor, *The Uralic Languages*, Routledge.

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in North Sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*. pages 66–75.

Miikka Silfverberg and Mans Hulden. 2018. Initial experiments in data-driven morphological analysis for Finnish. In *Proceedings of the Fourth International Workshop on Computatinal Linguistics of Uralic Languages*. pages 98–105.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2015. Finnpos: an open-source morphological tagging and lemmatization toolkit for finnish. *Language Resources and Evaluation* 50. https://doi.org/10.1007/s10579-015-9326-3.

Sindre Reino Trosterud, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto, and Kaisa Maliniemi. 2017. A morphological analyser for kven. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics, pages 76–88. https://doi.org/10.18653/v1/W17-0608.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*.