Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking

Tommi A Pirinen
Universität Hamburg
Hamburger Zentrum für Sprachkorpora
Max-Brauer-Allee 60, D-22765 Hamburg
tommi.antero.pirinen@uni-hamburg.de

Abstract

Building a treebank from scratch can easily be an elaborate, highly time consuming task, especially when working with a minority language with moderately complex morphology and no existing resources. It is also then typically true that language experts and informants with suitable skill sets are a very scarce resource. In this experiment I have attempted to work in parallel on building NLP resources while gathering and annotating the treebank. In particular, I aim to build a decent coverage morphologically annotated lexicon suitable for rule-based morphological analysis as well as accompanying rules for basic morphosyntactic analysis. I propose here a workflow, that I have found useful in avoiding redoing same work with related NLP resource construction.

1 Introduction

Karelian languages are languages closely related to Finnish spoken mainly in the republic of Karelia in Russia and surroundings. The languages are split in the ISO 639–3 standard between a few language codes: *Karelian* (krl) and *Livvi* or Olonets karelian (olo) for the two main branches of the language. The fact that 'krl' is commonly refered to as just Karelian can be confusing because 'olo' is also Karelian but I try to make the distinction clear throughout the article by using the ISO codes when necessary. The division is not totally unproblematic but I have followed it in the treebank for ease of development and use. There are some 35,000 native speakers of Karelian (krl) [1] and 31,000 for Livvi (olo) [2] according to Ethnologue, and both are classified as "Developing". The languages are developed enough to have some grammars (Zaikov, 2013; Ahtia, 1938; Markianova, 2002), dictionaries and books written, as well as some regular newspapers and broadcasts, but very few digital or computational resources so far. For unannotated corpora I have found a source with freely usable texts classified according to ISO language codes.

This paper discusses creation and ongoing work for two Karelian treebanks and compatible morphological parsers. The first part of the Karelian data will be included in the 2.4 release of the Universal Dependencies and I hope to enlarge and verify the data with native informants as well as include the Livvi data by the next release. The treebanks were named under the abbreviation of KKPP or *Karjalan kielten puupankit* which is Finnish for Karelian treebanks.

The rest of the article is organised as follows: in Section 2 I describe the languages and our goals for the treebanking, in Section 3 I describe the tools and methods for building treebanks, in Section 4 I describe the corpus selection and finally in Section 5 I summarise the article and talk about future work and ideas.

2 Background

As languages with very few available NLP resources, one of our first goals is to get annotated corpora. The universal dependencies format is a good choice for a standard for writing a new treebank at the moment; it has been used with many Uralic languages already that provide for reference for difficult

[1] https://www.ethnologue.com/18/language/krl/
[2] https://www.ethnologue.com/18/language/olo/

situations. Also, the North Saami treebank was made based on a rule-based finite-state morphological analyser (Sheyanova and Tyers, 2017), building one of which is also a goal for us, so I can safely say that the two formats are compatible and complement each other. One of the reasons why I make morphological analysers is to be able to provide number of end-user tools like spell-checking and correction as well as the reference corpus, for example in other Uralic languages there are plenty of resources hosted by giellatekno (Moshagen et al., 2014).

When I started with the treebanking, morphological analyser writing task, there were virtually no freely available corpora for Karelian and also no electronical dictionaries or analysers for Karelian krl. There was an existing analyser for Livvi and for that reason I have started our project with Karelian first. For digitised paper dictionaries, I have a dictionary for Karelian languages[3], that covers both Karelian and Livvi. The overall format and transcription differences, however, make it not directly usable for a source dictionary for morphological analyser for Karelian languages but rather an semi-automated source reference.

One of the thing I have established in the research of under-resourced languages in Uralic space is that for the survival and digital survival of a language certain technological resources need to be developed, and our aim with this project is to build as many of the necessary resources rapidly as possible.

One of the things that I have taken into consideration working on this treebank is how corpora are built within Uralic linguistic community outside the Universal Dependencies, e.g. in documentary linguistics. One of the prominent paradigms there is based on the line of tools from SIL shoebox to Fieldworks Explorer (FLeX), the workflow within those makes use of building corpora and dictionary simultaneously and this experiment is in a way our precursory study to implementing a similar tool for dependency treebanking style of linguistics. For reference on such Uralic research within computational linguistics see (Blokland et al., 2015).

Furthermore I are developing a morpho-syntactic rule-based methodology that can provide partial, ambiguous dependency graphs. The approach of building rule-based analysers first is very prominent within computational linguistics research of Uralic languages. In this article I are aiming to connect the traditional development of rule-based morphological analysers into treebanking workflow in a manner that optimises the usage of native informants' and the computational linguists' time, which is a crucial component for development in a very under-resourced setting.

Finally, I aim to have wide coverage of Uralic languages in the Universal Dependency project treebanks, and further study and experiment in the state-of-the-art methodology in large variety of NLP and typological research topics that have been empowered by the project. At the moment there are 6 Uralic treebanks available: Finnish (Haverinen et al., 2014; Voutilainen et al., 2012), Estonian (Muischnek et al., 2016), Hungarian (Vincze et al., 2010), North Saami (Sheyanova and Tyers, 2017), Komi (Partanen et al., 2018), and Erzya (Rueter and Tyers, 2018), out of some 30 that can easily have treebanks.

3 Methods

One of the contributions of this article is, that I am developing a sustainable workflow for creation of a wide array of technological resources for a seriously under-resourced language. For language technology infrastructure I will make use of an existing language technology infrastructure developed by (Moshagen et al., 2014), which I have selected because it provides a number of necessary components for free once morphological analysers are built, e.g. automatic spell-checking, machine-translation and so on.

The morphological analysis is based on the finite-state morphology (Beesley and Karttunen, 2003), this means in practice that one needs to build a dictionary and morphological rules describing the morphological processes. To couple the dictionary building with treebanking effort I have developed a method to generate lexicon entries from the annotated treebank data. I also use the analysers to generate suggestions for the annotators for the dependency annotations.

To give an example of the resource building workflow, a sentence might be annotated in CONLL-U format like:

```
# sent_id = vepkar-1774.7
```

[3] http://kaino.kotus.fi/cgi-bin/kks/kks_etusivu.cgi

```
# text = - Myö toivomma, jotta meijän kuccuh vaššatah starinankertojat ta guslinšoittajat, jotta kaccojat šuahah nähä
vanhanaikasien rahvahantapojen rekonstruointie, koroššetah järješteljät..
1       -       -       PUNCT   PUNCT   _       3       punct   _       Weight=0.0033333333333333335
2       Myö     myö     PRON    PRON    Case=Nom|Number=Sing|Person=1|PronType=Prs      3       nsubj   _       Weight
=500.0
3       toivomma        toivuo  VERB    VERB    Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 0
       root    _       Weight=0.0194|SpaceAfter=No
4       ,       ,       PUNCT   PUNCT   _       8       punct   _       Weight=518.6755555555555
5       jotta   jotta   SCONJ   SCONJ   _       8       mark    _       Weight=0.002142857142857143
6       meijän  myö     PRON    PRON    Case=Gen|Number=Plur|Person=1|PronType=Prs      7       nmod:poss       _
       Weight=500.0
7       kuccuh  kucču   NOUN    NOUN    Case=Ill|Number=Sing    8       obl     _       Weight=500.0
8       vaššatah        vaššata VERB    VERB    Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 3
       ccomp   _       Weight=500.0248
9       starinankertojat        starinan#kertoja        NOUN    NOUN Case=Nom|Number=Plur    15      nsubj   _  _
```

For a rule-based morphological parser an entry is needed to have at least dictionary form or lemma, and a paradigm for inflectional information; for languages like Karelian one cannot fully guess an entry for an inflectional paradigm from a single example but can usually give quite short list of plausible choices. So, I always extend our dictionaries with the entries from the annotated trees.

Likewise when annotating, I use the morphological analyser that is readily built with UD analyses: lemmas, UPOS and morphological features as well as some rough guesses when possible for the deps (e.g. puncts, Case-based dependencies); the python-based guesser for dependencies can currently handle things like: select `PUNCT` and suggest an attachment to each of the `VERB`s in sentence with `punct` dep, or select feature `Case=Acc` and suggest attachment to all `VerbForm=Fin` in the sentence with an `obj` dep. Thus, I can generate suggestion lists like:

```
# sent-id: <stdin>.21
# text: Koštamukšelaiset toivotah, jotta Koštamukšen ta Petroskoin šekä muijen
# kaupunkien välillä olis järješšetty šiännöllini lentoyhteyš.
1  Koštamukšelaiset  Koštamukšelaiset  X  X  _  _  _  _  SpaceBefore=No|_
2  toivotah  toivuo  VERB  VERBMood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act  0  root  _ _
2  toivotah  toivuo  VERB  Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Pass  0  root  _  SpaceAfter=No
3  ,  ,  SYM  SYM  _  _  _  _  SpaceBefore=No|Weight=506.4
3  ,  ,  PUNCT  PUNCT  _  2  punct  _  SpaceBefore=No|Weight=0.0033333333333333335
3  ,  ,  PUNCT  PUNCT  _  13  punct  _  SpaceBefore=No|Weight=0.03333333333333333
4  jotta  jotta  SCONJ  SCONJ  _  13  mark  _  Weight=0.0225
5  Koštamukšen  Koštamukšen  X  X  _  _  _  _  _
6  ta  ta  CCONJ  CCONJ  _  7  cc  _  Weight=0.01
7  Petroskoin  Petroskoi  PROPN  PROPN  Case=Gen|Number=Sing  2  obj  _  PropnType=Top|Weight=0.016666666666666666
```

A linguist is provided with this suggestion list per token in order defined by the weights,at the moment expert-determined rule-weighting but when we have large enough corpus I can easily incorporate the unigram log probabilities into weights as well. It should be noted that the linguist is allowed to discard all suggestions and this shall not be considered an unusual case while simultaneously building the analyser and the treebank. The current annotators also use an editor that is automatically running the validation tests[4] for UD after each edit and highlighting problems on the fly. The tools that I have developed so far will also be released with a free/libre open source licence.

When working on the annotation and guidelines I relied quite heavily on existing Uralic treebanks, especially Finnish since it is a closely related language with three treebanks and documentation. For many structures it is possible to find near or exact match using treebank search [5]. For example, the copula structure including the possession structure is marked in the same way in Finnish and Karelian languages, and generally many cases, function words and so forth, overlap with few systematic changes (e.g. in most parts of Karelian (krl) adessive and ablative have same form). Many of the examples where I did not find equivalents in Finnish I looked at other Uralic languages, or Russian, for example in elliptical structures a long hyphen is often used in Karelian and Russian to mark some elided tokens but not in contemporary Finnish in the genres of the UD treebanks at least.

Finally, this workflow goes on to ensure that the morphological analysers I build will have virtually a 100 % coverage of the treebank released, with a very high rate of recall for the treebank fields: lemma, UPOS and the lexical and morphological feature definitions. The reason recall is not 100 % is that there will be some annotations that, while theoretically correct, are not wanted in a normative analyser, e.g. colloquial uses of certain case forms in a role that is not the literary standard, as well as typos and mistakes,

[4] `https://github.com/universaldependencies/tools/validate.py`

[5] `http://bionlp-www.utu.fi/dep_search/`

Language	Lexicon size
Karelian	1452
Livvi	56,377

Table 1: The sizes of analysers of Uralic languages.

Treebank	Dependency trees	Syntactic words
Karelian	228	3094
Livvi	20	461
Finnish	34,859	377,822
Estonian	32,385	461,531
North Saami	3122	26,845
Hungarian	1800	42,032
Erzya	1550	15,790
Komi	307	3304

Table 2: The sizes of treebanks of Uralic languages. Dependency trees is number of annotated sentences and syntactic words as defined in UD guidelines.

however, I might change this practice in the future with universal feature `Style=Coll`.[6]

4 Data

There is not a great amount of available data written in Karelian languages to begin with. Furthermore, while there have been written texts for some time, the newest standard ortographies are quite recent, and there is some amount of variation from text to text in the written forms that is not the same as with older more standardised languages. Added to that is that telling languages apart, especially in less standard more dialectal writing, becomes non-trivial task. I started my data collection with web-crawling, and eventually found a corpus collection web site with open licencing policy, and the languages I want to work on categorised by language and genre, called VepKar.[7] The open licence also lets us work on articles instead of shuffled sentences, so it is another advantage.

By the time of writing I have developed a releasable treebank for Karelian and a morphological analyser, which are summarised in the table 2, I have also begun the work on Livvi treebank, which already had a usable analyser in place. For comparison I show some of the other existing Uralic treebanks for reference. Number of dependency trees annotated for non-Karelian languages is based on universaldependencies.org's statistics.

5 Discussion and future work

I have achieved a baseline universal dependency treebank and a morphological analyser for a minority language without pre-existing resources, and started working on a second treebank on a language with pre-existing analyser. In the next part I will contact more experts to verify the analyses and work on extending the treebanks as well as the analysers.

6 Acknowledgments

The author was employed by CLARIN-D during the project.

[6] I thank the anonymous reviewer for the helpful suggestion.

[7] `http://dictorpus.krc.karelia.ru/`

References

Edvard Vilhelm Ahtia. 1938. *Karjalan kielioppi*. Karjalan Kansalaisseura.

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI publications.

Rogier Blokland, Marina Fedina, Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2015. Language documentation meets language technology. In *Septentrio Conference Series*, number 2, pages 8–18.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.

Ludmila Markianova. 2002. Karjalan kielioppi 5-9. *Periodika, Petroskoi*.

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 e Ninth International Conference on Language Resources and Evaluation, LREC*, pages 71–77.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian dependency treebank: from constraint grammar tagset to universal dependencies. In *LREC*.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first komi-zyrian universal dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132.

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 106–118.

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.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank.

Atro Voutilainen, Kristiina Muhonen, Tanja Katariina Purtonen, Krister Lindén, et al. 2012. Specifying treebanks, outsourcing parsebanks: Finntreebank 3. In *Proceedings of LREC 2012 8th ELRA Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Pekka Zaikov. 2013. *Vienankarjalan kielioppi*.