Endangered Languages are not Low-Resourced!

Mika Hämäläinen[0000−0001−9315−1278]

University of Helsinki, Finland
mika.hamalainen@helsinki.fi

Abstract. The term low-resourced has been tossed around in the field of natural language processing to a degree that almost any language that is not English can be called "low-resourced"; sometimes even just for the sake of making a mundane or mediocre paper appear more interesting and insightful. In a field where English is a synonym for language and low-resourced is a synonym for anything not English, calling endangered languages low-resourced is a bit of an overstatement. In this paper, I inspect the relation of the endangered with the low-resourced from my own experiences.

Keywords: endangered languages · low-resourced languages · NLP · criticism of science

1 Introduction

Low-resourced[1] or *low-resource* language is one of those terms that has never been defined in our field, and yet it has been used very often in many publications in the past, present and future. As it is a term that is supposed to be implicitly understood without any actual thresholds for the number of speakers or the amount of data etc., it is no wonder that people use the term as they please.

Several tasks have been called low-resourced in the NLP research, for languages such as: Chinese [40] (1.2 billion speakers), Arabic [9] (422 million speakers), Bengali [24] (228 million speakers), Japanese [41] (126 million speakers), Vietnamese [16] (76 million speakers), Dutch [22] (24 million speakers), Sinhala [15] (17 million speakers) and Finnish [2] (5 million speakers). To make matters worse, even endangered languages have been called low-resourced [14,38,20,18]. The listing is not exhaustive, but it illustrates the problem how little semantic value the term low-resourced has, given that any language can be called low-resourced.

As someone who is native in a relatively small language, Finnish, I have a very different perspective to what low-resourced means. Finnish has around 5 million speakers, and since that is around a half of the population of Tokyo (9.3 million), our language does seem small. At any case, I would never call my native language low-resourced[2]. Why is this, you might ask? Calling Finnish

[1] The form preferred by Dr Jack Rueter

[2] Unless I wanted to get a mediocre paper accepted

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 1–11. ––

low-resourced is denying the fact that we have our own TV shows, movies, music, theater plays, novels and other cultural products in Finnish. And I am not even talking about small numbers here. Besides, Finnish is a language of education, there is no level of schooling you could not complete entirely in Finnish, from preschool to defending your PhD. Yes, our language is small on a global scale, but we do have a whole bunch of resources we generate every day by communicating with each other!

Now, one of the main problems I find with the term low-resourced, when used about languages that have millions of speakers, is that the resources are always out there. It is, perhaps, more often than not a question of learned helplessness of a researcher. There are many ways of doing annotation projection or just gathering data in a savvy way by crawling the internet. If some NLP resource does not exist for a language, stop complaining about how low-resourced it is, get up and gather the data. Of course, there are always exceptions when gathering the data required for a large language might not be a walk in a park such as when dealing with historical data [11]. And it is true that even resources for non-endangered languages can be noisy [19]. However, working with a non-endangered language does not have the same degree of problems as endangered languages might have, as I will describe later in this paper.

Quite often there are a lot of annotated resources for a majority language that are hidden on a hard drive of a researcher or published somewhere where they can be difficult to find. This leads to a situation where a language might seem low-resourced initially, but the resources are already out there, pre-annotated. They are just hidden somewhere where Google will not find them, or in the Nordic context, hidden behind a Korp user interface [4] in such a way that the resources have no value for NLP.

The main purpose of this paper is to wake people working with NLP up. If we want to continue using the term low-resourced, we had better define it as a community. Or much rather come up with a classification of how low-resourced a language is. It is about time we stopped using the term low-resourced as a fancy term to boost our papers or as an excuse for not annotating data (and releasing it for others to use).

2 Endangered, but How Endangered?

For the reasons that should have become evident by now, I feel that calling any endangered language low-resourced is an overstatement, as it clearly lifts a truly resource-poor language into the same league with the big players. Before continuing any further in this section, I would like to point out that much like low-resourced languages, endangered languages are not a homogeneous group. There is a huge variety in the digital resources and socio-political status these languages and their speakers have. As an example, UNESCO [21] categorizes endangered languages into 5 categories: vulnerable, definitely endangered, severely endangered, critically endangered and extinct, depending on the level of intergenerational language transmission.

In this section, I will describe some of my encounters with endangered language communities. I know that my experiences do not reflect all endangered languages, but I feel that it is important to share them to better contextualize what endangered languages can be. This is something that gets easily forgotten when doing NLP research as a rich language and culture get very easily reduced into a dataset, machine learning model and results.

I had an opportunity to visit Ufa, the capital of the Republic of Bashkortostan, Russia. The local language, known as Bashkir (bak), is rated vulnerable according to UNESCO with its more than 1.6 million speakers. While visiting the premises of Bashkir Encyclopedia[3] (*Башкирская энциклопедия*), it became evident to me that there was no lack of high-quality written knowledge in Bashkir. The number of different encyclopedias about different topics they showed us was incredible. Not to mention that they were very serious about writing the encyclopedias, according to them, they only believed in facts and numbers. The encyclopedias were not thus just mere translations of existing ones, but rather their own independent work of science. I was also told about TV channels broadcasting TV shows in Bashkir, which while interesting, is not that surprising given the number of speakers.

FU-Lab[4] welcomed me for a research visit in the capital of the Komi Republic, Russia, Syktyvkar. While there are several Komi languages, our discussions were mainly related to Komi-Zyrian (kpv), a language marked as definitely endangered in the classification of UNESCO with as few as 217000 native speakers. Seeing the work conducted in FU-Lab in action, one can say that Komi-Zyrian has a surprising amount of digital resources. They actively develop constraint grammar based disambiguators and contribute to the morphological FSTs (finite-state transducers). In addition to that they compile text and audio corpora for Komi (see [7,6,5]) and have many dictionaries available in a digital format. An interesting decision by the Komi Republic that could eventually lead into functional Komi-Russian machine translation is the fact that all Russian law texts are translated into Komi. At FU-Lab, they have ensured that the translations remain parallel to the original Russian law texts, which should make machine translation easy in the future.

Our system Ve′rdd [1] was the reason I got an opportunity to visit the Sami Culture Center Sajos[5] in Inari, Finland to collaborate with two Skolt Sami dictionary editors. Skolt Sami (sms) is a severely endangered language with only 300 native speakers according to UNESCO. Despite the low number of speakers, they had the presentations of the Sami cultural event simultaneously interpreted from Skolt Sami to Finnish and from other Sami languages to Skolt Sami by professional interprets. Thanks to Rueter's continuous efforts for the digital revitalization of the language, Skolt Sami has an extensive digital multilingual dictionary [30] and FST morphology [27]. The situtaion of Skolt Sami is fortunate in the sense that it is one of many Sami languages. The Sami Parliament

[3] http://www.bashenc.ru/

[4] https://fu-lab.ru/

[5] http://www.sajos.fi/

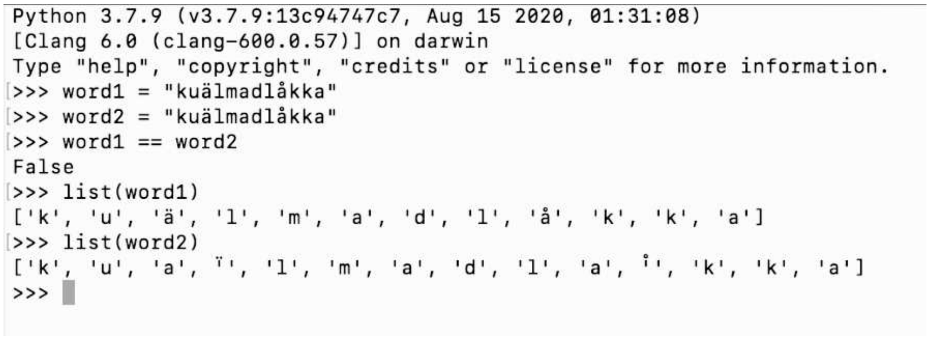
has established Sámi Giellagáldu to do work on language norms and terminology for various Sami languages including Skolt Sami.

As a summary, it is important to understand that the term endangered language is complicated as well in terms of the linguistic resources available for NLP tasks. Some endagered languages may have a surprising amount of resources for some specific NLP tasks, while others may not have digital resources at all. For more anecdotes, I strongly recommend reading Rueter's personal experiences of everyday situations in Erzya [26].

3 The Underappreciated Problem of Being Endangered

There have been several papers trying to tackle low-resourced tasks either by simulating a resource-poor scenario in a high-resourced language or by having limited resources for a high-resourced language [8,39,13]. It is important to remember that while these efforts are of value in many domains, they might not be directly applicable as such for endangered languages.

The issues you might face with an endangered language start from the very low-level: character encoding. I am not referring to any custom or local encoding, but to Unicode, the encoding we know and love. Unicode is not at all as uncomplicated when we are dealing with smaller languages. One of the problems is that Unicode has multiple ways of encoding one character and that there are similar looking, but not quite the same characters.



```
Python 3.7.9 (v3.7.9:13c94747c7, Aug 15 2020, 01:31:08)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> word1 = "kuälmadlåkka"
[>>> word2 = "kuälmadlåkka"
[>>> word1 == word2
False
[>>> list(word1)
['k', 'u', 'ä', 'l', 'm', 'a', 'd', 'l', 'å', 'k', 'k', 'a']
[>>> list(word2)
['k', 'u', 'a', '̈', 'l', 'm', 'a', 'd', 'l', 'a', '̊', 'k', 'k', 'a']
>>> 
```

Fig. 1. A typical situation with Skolt Sami data.

Figure 1 illustrates the aforementioned problem. The word *kuälmadlåkka* seems to be written similarly in *word1* and *word2*. However, they are not identical, as seen when they are split into characters. This type of an issue is not as common in high-resourced languages and it might go unnoticed for the unwary researcher. These issues are present with some endangered languages because of a multitude of reasons such as a simple lack of a suitable keyboard layout, lack of a standardized orthography or a change in orthography. Although, for practical reasons, you might see people writing words in a non-standard way even in larger languages such as writing *paral.lel* instead of *paral·lel* in Catalan or *ca* instead of *ça* in French, however the reason for these non-standard ways of writing are different than in the Skolt Sami example. At any rate, misspellings are common in texts written in an endangered language as well, as reported for North Sami [3].

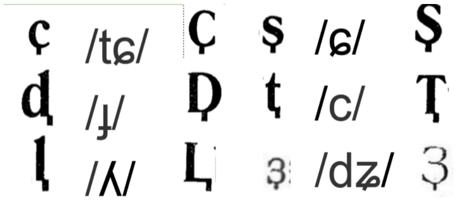


Fig. 2. Latinitsa letters missing from Unicode

Sometimes not all important characters are in Unicode as seen in Figure 2. Such is the case with the writing system of Komi-Zyrian that was in place before the Cyrillics, латиница (latinitsa). There has been a long-running effort [34] in getting the missing characters into the Unicode standard, but so far this effort has been futile. This means that there is no consistent way of encoding historical Komi texts written in this script.

When we write anything in a big language, it is usually clear what is right and wrong. An extensive part of any educational system goes into teaching people to write in a normative way. Much of this normative writing is something we who are native in a non-endangered language just learn by being exposed to normative language in various places such as books and news. Norms are usually established and maintained by a language institute. In Finland, this practice is called *kielenhuolto* (language maintenance), a metaphor that makes any non-normative language sound broken like any machine that is in need for maintenance. Even though the "correct" way of using a language is internalized to different levels by different people, this practice will be reflected in any data produced in a high-resourced language.

The tradition of seeking for a correct language is called prescriptive linguistics, where the focus is on how a language should be. Sometimes this prescriptive ideology is absent or the prescriptive rules are not internalized by the people native in an endangered language due to various reasons such as a limited access to education in one's own language. This leads to a situation where we cannot expect endangered language data to be "correct" in the prescriptive sense, but it might be more reflective of how people actually use the language. This still does not mean that the data would be consistently deviant of some norm. Just as with bigger languages, endangered languages, even as small as Skolt Sami have different dialects [31].

With the recent edition of the Skolt Sami dictionary [17] there was some discussion about certain words such as why *Amerikk* (America) and *ankerias* (eel) were used instead of older words *Ä′mmrikk* and *aŋŋerias*. Questions like these are related to whether a language should be documented as it should be or as it is. A highly prescriptive dataset in an endangered language might thus mean no applicability to real world data as people using the language might not be aware of the norms. In the case of Skolt Sami, I would be surprised if all speakers were aware of the rules, as Sámi Giellagáldu publishes their latest recommendations in their blog, making such recommendations difficult to consult.

Endangered language data is also more prone to containing mistakes beyond encoding and lack of norms. In my work, I have noticed several mistakes in different resources such as XML dictionaries [14]. I would not point fingers and call anyone's work bad, as mistakes do happen, especially when there are several people working on the resources during different times. The reason why I believe that there are more mistakes in endangered language resources is the simple fact that there are fewer people inspecting them and pointing out errors. It is very common to become blind to one's own work, and spotting mistakes requires external inspection.

4 Do Only Rules Rule?

One of the things that divide people in NLP is rules versus neural networks. Why would you write rules for an endangered language if neural networks work for a low-resourced Hindi? As we have seen in this paper, the problems endangered languages have are not the same as just about any "low-resourced" language would have. But at the same time I am facing the ideology that only rules can be used to model endangered languages. I, myself, don't believe that either rules or neural networks are the answer. The optimal solution is probably somewhere in the middle ground.

Rule-based methods are continuously developed for various endangered and extinct languages [37,23,36]. Sometimes, due to the lack of resources, rules are the only viable way of dealing with these languages. I believe that here rules serve for a more important role than mere engineering of an NLP system. Many endangered languages are under-documented, and machine readable rules serve for language documentation purposes as well. They need to capture something meaningful about the language being described in order for them to work to begin with. From this perspective, I think that rule-based systems are not only valuable from the point of view of NLP but also from the point of view of linguistic research. Only machine readable rules let you test out your linguistic hypotheses extensively.

Rules are good in the sense that they can be fixed easily, and it is possible to reach to a high accuracy with them. This is useful when building systems like spell checkers and language learning tools, as the correctness of these tools is of utmost importance. However, rules can only go so far. An FST, no matter how extensive, is never going to contain all the words in its vocabulary, for example.

For this reason, neural networks are useful, as they can produce output even for new input that was not present in the training data. However, usually accuracy is important in the context of endangered languages as many of the NLP tools are built for practical purposes and for the benefit of the language community.

Rules can get things right, but their limits can be reached easily. Neural networks can go beyond a predefined set of rules, but they are more prone to producing incorrect results as well. I think that the two different approaches should be used together. Rules can be used to generate training data for neural networks, something I call *fake it till you make it approach*, or they can be used to filter out low-quality samples from a training dataset. This way, rules can be used in the training process. Because rules can be easily fixed, I would pipeline rules with neural networks. Whatever rules cannot cover, a neural network can handle, and if the system produces wrong output, the rule-based method can always be fixed.

I think that synthetic data generation is still an under-studied way of building NLP tools for endangered languages. FSTs are a good way of achieving this as they can be used both for generation and analysis. We have reached to rather good results for some languages with FST generated data and you can expect to see neural models integrated with UralicNLP [12][6] as a backup for failing FSTs in the near future.

5 Conclusions

The term low-resourced is truly a complicated one and it makes NLP research conducted for endangered languages difficult to get published in the bigger ACL venues. Work with endangered languages is not as state-of-the-art driven as it is usually to be expected from NLP papers in bigger venues. Instead such a work is more practical, typically involving producing tools and resources for the benefit of the language community.

Any work with endangered languages includes ethical considerations. I have always been puzzled by the fact that in the world of NLP research not releasing one's code, data and models is considered acceptable practice. With endangered languages not releasing the resources produced is even more severe as such a behavior may be interpreted more as a cultural and linguistic appropriation of a vulnerable group of people purely for the sake of academic merit.

Endangered languages pose very different types of challenges for NLP research and they have very different amounts and types of resources available. Some languages have quite advanced NLP tools in place thanks to altruistic research endeavors and active community members, while others do not have anything. As there is such a huge variation within the group of endangered languages, grouping them together with anything "low-resourced" from Chinese to Finnish[7] is very misleading.

[6] https://github.com/mikahama/uralicNLP

[7] I still don't think Finnish is low-resourced

I have shared my personal experiences working with NLP for endangered languages and working with people who are native in some of them. A lot has been left unsaid, and I do know that there are a whole lot of languages out there that are dealing with very different issues than what I have described in this paper (c.f. [10]). The main purpose of my descriptions has been to show people what type of problems one can encounter when conducing this type of a research. I am honored for having had this possibility of seeing NLP beyond large languages.

Acknowledgments

I would like to thank Jack Rueter, who was kind enough to supervise my PhD thesis even though it ended up being unrelated to Jack's research interests. Without him, I would have never been exposed to endangered languages nor would have I ever had the chance of traveling to the places where such languages are spoken. Also, without his fantastic resources like FSTs [28,25], dictionaries [29], corpora [32] and Universal Dependencies treebanks [35,33] NLP would be a distant dream for many endangered Uralic languages.

References

1. Alnajjar, K., Hämäläinen, M., Rueter, J., Partanen, N.: Ve'rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. In: Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations. pp. 1–6 (2020)
2. Alnajjar, K., Leppänen, L., Toivonen, H.: No time like the present: methods for generating colourful and factual multilingual news headlines. In: Proceedings of the 10th International Conference on Computational Creativity. Association for Computational Creativity (2019)
3. Antonsen, L.: Cállinmeattáhusaid guorran.[english summary: Tracking misspellings.]. University of Tromsø (2013)
4. Borin, L., Forsberg, M., Roxendal, J.: Korp-the corpus infrastructure of språkbanken. In: LREC. pp. 474–478 (2012)
5. Федина, МС: Создание официально-делового подкорпуса национального корпуса коми языка. In: УПРАВЛЕНИЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМ РАЗВИТИЕМ СУБЪЕКТА РОССИЙСКОЙ ФЕДЕРАЦИИ, pp. 205–216 (2015)
6. Федина, МС and Левченко, ДА: Из опыта создания коми медиатеки. In: ЭЛЕКТРОННАЯ ПИСЬМЕННОСТЬ НАРОДОВ РОССИЙСКОЙ ФЕДЕРАЦИИ: ОПЫТ, ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ. pp. 220–227 (2017)
7. Федина, Марина Серафимовна: Корпус Коми Языка Как База Для Научных Исследований. In: II Международная научная конференция «Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы» проводится в рамках реализации Государственной программы «Сохранение и развитие государственных языков Республики Башкортостан и языков народов Республики Башкортостан» на 2019–2024 гг. Ответственный редактор: Ахмадеева АУ. p. 45 (2019)

8. Gu, J., Hassan, H., Devlin, J., Li, V.O.: Universal neural machine translation for extremely low resource languages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 344–354 (2018)
9. Halabi, D., Awajan, A., Fayyoumi, E.: Arabic lfg-inspired dependency treebank. In: 2017 International Conference on New Trends in Computing Sciences (ICTCS). pp. 207–215. IEEE (2017)
10. Hammarström, H.: The status of the least documented language families in the world. Language Documentation & Conservation **4**, 177–212 (2010)
11. Hill, M.J., Hengchen, S.: Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. Digital Scholarship in the Humanities **34**(4), 825–843 (04 2019). https://doi.org/10.1093/llc/fqz024, https://doi.org/10.1093/llc/fqz024
12. Hämäläinen, M.: UralicNLP: An NLP library for Uralic languages. Journal of Open Source Software **4**(37), 1345 (2019). https://doi.org/10.21105/joss.01345
13. Hämäläinen, M., Alnajjar, K.: A template based approach for training nmt for low-resource uralic languages - a pilot with finnish. In: ACAI 2019: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. pp. 520–525. ACM, United States (Dec 2019). https://doi.org/10.1145/3377713.3377801
14. Hämäläinen, M., Tarvainen, L., Rueter, J.: Combining concepts and their translations from structured dictionaries of uralic minority languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 862–867 (2018)
15. Karunanayake, Y., Thayasivam, U., Ranathunga, S.: Transfer learning based free-form speech command classification for low-resource languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 288–294. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-2040, https://www.aclweb.org/anthology/P19-2040
16. Kruengkrai, C., Nguyen, T.H., Aljunied, S.M., Bing, L.: Improving low-resource named entity recognition using joint sentence and token labeling. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5898–5905. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.523, https://www.aclweb.org/anthology/2020.acl-main.523
17. Lehtinen, M., Koponen, E., Fofonoff, M., Lehtola, R., Rueter, J. (eds.): Suomi–koltansaame-sanakirja Lääʹdd-sääʹm-sääʹnnǩeeʹrjj. Saamelaiskäräjät (2021)
18. Lim, K., Partanen, N., Poibeau, T.: Multilingual dependency parsing for low-resource languages: Case studies on north saami and komi-zyrian. In: Language Resource and Evaluation Conference (2018)
19. Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., Nevalainen, T.: Wrangling with non-standard data. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 81–96 (2020)
20. Moeller, S., Kazeminejad, G., Cowell, A., Hulden, M.: Improving low-resource morphological learning with intermediate forms from finite state transducers. In: Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers). pp. 81–86. Association for Computational Linguistics, Honolulu (Feb 2019), https://www.aclweb.org/anthology/W19-6011

21. Moseley, C. (ed.): Atlas of the World′s Languages in Danger. UNESCO Publishing, 3rd edn. (2010), online version: http://www.unesco.org/languages-atlas/
22. Perl, T., Chaudhury, S., Giryes, R.: Low resource sequence tagging using sentence reconstruction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2692–2698. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.239, https://www.aclweb.org/anthology/2020.acl-main.239
23. Pirinen, T.A.: Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in karelian treebanking. In: Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019). pp. 132–136 (2019)
24. Rezaul Karim, M., Kanti Dey, S., Raja Chakravarthi, B.: Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. arXiv e-prints pp. arXiv–2012 (2020)
25. Rueter, J.: Хельсинкиса университетын кыв туялысь Ижкарын перымса кывъяс симпозиум вылын лыддьӧмтор. Permistika pp. 154–158 (2000)
26. Rueter, J.: The erzya language. where is it spoken? Études finno-ougriennes (45) (2013)
27. Rueter, J., Hämäläinen, M.: Fst morphology for the endangered skolt sami language. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). pp. 250–257 (2020)
28. Rueter, J., Hämäläinen, M., Partanen, N.: Open-source morphology for endangered mordvinic languages. In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS). pp. 94–100. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.nlposs-1.13, https://www.aclweb.org/anthology/2020.nlposs-1.13
29. Rueter, J., Hämäläinen, M., et al.: On xml-mediawiki resources, endangered languages and tei compatibility, multilingual dictionaries for endangered languages. In: AsiaLex 2019 Proceedings of the 13th Conference of the Asian Association for Lexicography. Asos Publisher (2019)
30. Rueter, J., Hämäläinen, M.: Synchronized mediawiki based analyzer dictionary development. In: Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages. pp. 1–7 (2017)
31. Rueter, J., Hämäläinen, M.: Skolt Sami, the makings of a pluricentric language, where does it stand?, pp. 201–208. No. 21 in Österreichisches Deutsch – Sprache der Gegenwart, Peter Lang (2020)
32. Rueter, J., Partanen, N.: On new text corpora for minority languages on the helsinki korp.csc.fi server. In: Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы. p. 32–36 (2019)
33. Rueter, J., Partanen, N., Ponomareva, L.: On the questions in developing computational infrastructure for Komi-permyak. In: Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages. pp. 15–25. Association for Computational Linguistics, Wien, Austria (10–11 Jan 2020), https://www.aclweb.org/anthology/2020.iwclul-1.3
34. Rueter, J., Ponomareva, L.: Komi latin letters, degrees of unicode facilitation. In: Proceedings of the Language Technologies for All (LT4All) (2019)
35. Rueter, J., Tyers, F.: Towards an open-source universal-dependency treebank for Erzya. In: Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. pp. 106–118. Association for Computational

Linguistics, Helsinki, Finland (Jan 2018). https://doi.org/10.18653/v1/W18-0210, https://www.aclweb.org/anthology/W18-0210

36. Sahala, A., Silfverberg, M., Arppe, A., Lindén, K.: BabyFST - towards a finite-state based computational model of ancient babylonian. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 3886–3894. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.479
37. Schmirler, K., Arppe, A.: Modelling plains cree negation with constraint grammar. In: Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar-Methods, Tools and Applications, 30 September 2019, Turku, Finland. pp. 27–34. No. 168, Linköping University Electronic Press (2019)
38. Soisalon-Soininen, E., Granroth-Wilding, M.: Cross-family similarity learning for cognate identification in low-resource languages. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 1121–1130. INCOMA Ltd., Varna, Bulgaria (Sep 2019), https://www.aclweb.org/anthology/R19-1129
39. Tiedemann, J.: The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. In: Proceedings of the Fifth Conference on Machine Translation. pp. 1174–1182 (2020)
40. Xu, J., Ma, S., Zhang, Y., Wei, B., Cai, X., Sun, X.: Transfer deep learning for low-resource chinese word segmentation with a novel neural network. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 721–730. Springer (2017)
41. Yu, C., Han, J., Zhang, H., Ng, W.: Hypernymy detection for low-resource languages via meta learning. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3651–3656. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.336, https://www.aclweb.org/anthology/2020.acl-main.336