Uralic Language Identification (ULI) 2020 shared task dataset and the Wanca 2017 corpora

Tommi Jauhiainen
Department of Digital Humanities
University of Helsinki
tommi.jauhiainen@helsinki.fi

Heidi Jauhiainen
Department of Digital Humanities
University of Helsinki
heidi.jauhiainen@helsinki.fi

Niko Partanen
Department of Finnish, Finno-Ugrian
and Scandinavian Studies
University of Helsinki
niko.partanen@helsinki.fi

Krister Lindén
Department of Digital Humanities
University of Helsinki
krister.linden@helsinki.fi

Abstract

This article introduces the Wanca 2017 web corpora from which the sentences written in minor Uralic languages were collected for the test set of the Uralic Language Identification (ULI) 2020 shared task. We describe the ULI shared task and how the test set was constructed using the Wanca 2017 corpora and texts in different languages from the Leipzig corpora collection. We also provide the results of a baseline language identification experiment conducted using the ULI 2020 dataset.

1 Introduction

As part of the Finno-Ugric Languages and the Internet project, (SUKI)[1] we have collected textual material for some of the more endangered Uralic languages from the Internet (Jauhiainen et al., 2015a). In this paper, we introduce the Wanca 2017 corpora which will be published in the Language Bank of Finland[2] as a downloadable package as well as through the Korp[3] concordance service. The Uralic Language Identification (ULI) 2020 shared task[4] was organized as part of the VarDial 2020 Evaluation campaign.[5] In order to create a training dataset for the shared task, we used the earlier version of the corpora, Wanca 2016[6] (Jauhiainen et al., 2019a), together with corpora available from the Leipzig corpora collection[7] (Goldhahn et al., 2012). Different corpora from the Leipzig corpora collection and a manually verified subset of the Wanca 2017 corpora were used to create the test set for the shared task. We also performed a baseline language identification experiment for the ULI dataset using the HeLI method described by Jauhiainen et al. (2017b).

In this paper, we first introduce some related work and resources for language identification and the Uralic languages in Section 2. We then describe the Wanca 2017 corpora and its creation in Section 3.

[1] http://www.suki.ling.helsinki.fi/eng/project.html

[2] https://www.kielipankki.fi/language-bank/

[3] https://korp.csc.fi

[4] https://sites.google.com/view/vardial2020/evaluation-campaign/uli-shared-task

[5] https://sites.google.com/view/vardial2020/evaluation-campaign

[6] http://urn.fi/urn:nbn:fi:lb-2020022901

[7] https://corpora.uni-leipzig.de

In Section 4, we give a detailed description of the creation of the dataset for the ULI 2020 shared task as well as the information about the baseline language identification experiments using the dataset. Finally, we provide some error analysis for the results of those experiments.

2 Related work

In this section, we first introduce some previous work on language identification of texts, then we give a short introduction to the Uralic languages and present some of the text corpora already available for those languages.

2.1 Language identification in texts

In this paper, we focus on language identification in texts as opposed to language identification in speech. By language identification, we mean the labeling of sentences or texts by language labels from a given label set, which is the test set-up in the ULI shared task. By defining the problem this way, we have ignored two challenges in language identification: detection of unknown languages and handling multilingual texts. In unknown language detection, the language identifier can be presented with texts that are written in a language that it has not been trained in. Multilingual texts contain parts written in more than one language. Actually, in the strict sense, some of the sentences in the training and test sets of the ULI task can be considered multilingual as they may include some words in languages other than the main language of the sentence. These kind of multilingual sentences are present especially in the corpora for the non-relevant languages. In the ULI task, the target is however, to simply label the main language for each sentence.

A recent survey concerning language identification in texts by Jauhiainen et al. (2019b) gives a thorough introduction to the subject.

2.2 Uralic languages

In this section we provide a general overview to the Uralic language family, with specific attention to development of the written standards and contemporary use, as this is closely connected to the resources available for the language identification task. The Uralic language family contains 30-40 languages, and shows considerable diversity at all levels. Handbooks about the family include Abondolo (1998) and Sinor (1988), and new handbooks are currently under preparation (Bakró-Nagy et al., forthcoming; Abondolo and Valijärvi, forthcoming). The Uralic language family is one of the most reliably established old language families in the world, and can be compared with the Indo-European language family in its time depth and variation, although the exact dating of the family is a matter of on-going research.

Geographically, the Uralic languages are spoken in Northern Eurasia, with the Saami languages in the Scandinavia representing the westernmost extent, and the Nganasans at the Taimyr Peninsula being the easternmost Uralic language speakers. In the south, Hungarian, a geographical outlier, is spoken in the Central European Carpathian Basin. The majority of the Uralic languages are spoken within the Russian Federation. The wide geographical area also has resulted in different subsistence strategies and livelihoods, historically, and also in various contemporary conditions. Only three Uralic languages, Estonian, Finnish and Hungarian, are spoken as the majority language of a country. These languages are not endangered, but they have closely related varieties that often are endangered, as are all other Uralic languages.

Some Uralic languages are already extinct. This is the case with Kemi Saami, which ceased to be spoken in the 19th century, and Kamas, the last speaker of which died in 1989. The former is represented in this shared task as texts written in it were found from the Internet and they are now part of the Wanca 2016 corpora. Still spoken Uralic languages form a continuum also in their number of speakers, as the smallest languages, such as Inari Saami and Skolt Saami, have only hundreds of speakers, and Nganasan maybe slightly more than one hundred (Wagner-Nagy, 2018, 17). On the contrary, languages such as Mari or Udmurt have hundreds of thousands of speakers, and are used actively in various spheres of modern society. They are, nevertheless, endangered due to interrupted intergenerational language transmission and disruption of the traditional speech communities.

When it comes to the online presence, or generally to available textual representations of these languages, historical developments in their standardization and language planning play a very central role. This was largely outlined by Soviet language policy, described in detail in Grenoble (2003). It has also been typical for the Uralic languages spoken in Russia that their orthographies have changed numerous times. Siegl and Rießler (2015) discuss four case studies about the possible variation in the degrees of contemporary literacy and development of the written standards. There are numerous languages in the Wanca corpora for which the ortographies were developed and repeatedly changed in the late 19th or early 20th century. This pertains especially to many languages spoken in the Soviet Union, including Ingrian, Karelian, Livvi-Karelian, Vepsian, Komi-Permyak, Komi-Zyrian, Udmurt, Khanty, Mansi and Tundra Nenets. Even with very closely related languages the contemporary orthographies and the language varieties themselves contain numerous differences in their phonology and spelling conventions that make distinguishing the language of a text almost always straightforward, at least to a specialist. Such closely related languages with clearly distinct written traditions include two Komi written standards, two Mari written standards and two Mordva written standards. These differences are large enough that, from the perspective of computational linguistics, distinct infrastructure usually has to be developed for each variety, even if the actual linguistic differences would be minor. For an example of challenges in creating an infrastructure for Komi-Permyak and Komi-Zyrian see Rueter et al. (2020) and for Mordvinic languages see Rueter et al. (in press).

Some of these orthographies were more successful than others, and there is large variation in when exactly the currently used systems were established and what level of stability they have. For example, the orthography created in 1986 for Nganasan was never widely used, and in the current orthography the conventions vary with author and editor (Wagner-Nagy, 2018). For languages such as Votic, the current orthography was developed first in the 2000s (Èrnits, 2006, 3). An earlier example is Tundra Nenets, which has had the current orthography since the 1940s, and which has all in all 100 titles published. The language is also partially used in local newspapers (Nikolaeva, 2014). However, the small number of Tundra Nenets sentences in the Wanca corpora probably indicates that the online visibility of the language is relatively small. At the same time a relatively small Saami language, Skolt Saami with approximately 300 speakers, is represented in the dataset by thousands of sentences. The Skolt Saami orthography was developed in the 1970s and the knowledge of the writing standard has not reached the whole community (Feist, 2015, 26,37), but the language has been officially recognized in Finland and has received state support, which may explain why it appears to have more online presence than some other languages of the same size.

The majority of the Uralic languages spoken in Russia are nowadays written with Cyrillic orthographies. Exact orthographic conventions differ from Russian, but similar conventions are regularly employed, i.e. to express palatal or palatalized phoneme distinctions. Some languages, such as Erzya, have essentially the same character set as Russian, whereas most of the languages have additional characters. Some of these are shared by numerous languages that use Cyrillic orthography, such as *Cyrillic O with diaeresis*, which is used in Komi, Mari and Udmurt orthographies. There is also the example of *Ze with diaeresis*, which is used only in the Udmurt orthography. For the language identification task these characters can be very valuable cues about the language, but as they are not necessarily present in all keyboards, online texts are also regularly found where they are replaced with other characters or conventions. Finnic languages spoken in Russia are written with Latin orthographies, although historically also some Cyrillic orthographies have been in use.

Thereby the contemporary online presence of the Uralic languages is a complex combination of many historical factors. However, we can generally say that the languages with more widely used and taught orthographies, and with a substantial speaker base, do have enough materials online that downloading up to several million tokens is possible. With smaller languages the situation is different and much more varying. There is also the aspect of time, as continuous use accumulates increasingly larger resources. When it comes to extinct languages, their corpora have to be considered finite.

2.3 Corpora for Uralic languages

For the Uralic languages that are the majority language of a country, that is Finnish, Estonian, and Hungarian, many large text corpora already exist. For example, there is the Suomi 24 Corpus[8] with over 250 million Finnish sentences from a social networking website available from the Language Bank of Finland, and the Europarl corpus[9] with over 600,000 sentences of Hungarian and Estonian (Koehn, 2005). The Leipzig Corpora Collection[10] has texts also for some of the more rare Uralic languages: Eastern Mari, Komi, Komi-Permyak, Northern Saami, Udmurt, Võro, and Western Mari. The Giellatekno research group has three Korp installations for Uralic languages: one[11] for Saami languages, one[12] for Kven, Meänkieli, Veps, and Võro, and one[13] for Komi-Zyrian, Komi-Permyak, Udmurt, Moksha, Erzya, Hill Mari, and Meadow Mari. The Wanca in Korp corpora contain texts in all the aforementioned languages as well as some additional Uralic languages.[14] Several endangered Uralic languages also have treebanks in the Universal Dependencies project.[15] These include Northern Saami (Tyers and Sheyanova, 2017), Komi-Zyrian (Partanen et al., 2018), Komi-Permyak, Erzya (Rueter and Tyers, 2018), Moksha, and two Karelian varieties (Pirinen, 2019). Under construction in the Language Bank of Finland is also the Parallel Bible Verses for Uralic Studies corpus (PaBiVus), which contains Bible translations from different publications.[16]

Especially in the context of this shared task it is important to mention previous work that has collected online texts in the minority languages spoken in Russia. At least Orekhov et al. (2016) and Krylova et al. (2015) have collected online and social media texts in various languages, and Arkhangelskiy (2019) has published corpora of this type for Uralic languages. Wanca 2017 corpora, described next, connects well to the earlier work.

3 Wanca 2017 corpora

The aim of the SUKI project was to find texts written in Uralic minority languages from the Internet (Jauhiainen et al., 2015a). The set of relevant languages was determined as all the Uralic languages included in the ISO 639-3 standard except Finnish, Estonian, and Hungarian. In order to find the texts, we used an open-source web-crawler called Heritrix (Mohr et al., 2004) combined with different language identifiers we were developing during the project (Jauhiainen et al., 2015b; Jauhiainen et al., 2015c; Jauhiainen et al., 2016; Jauhiainen et al., 2017a; Jauhiainen et al., 2017b). In addition to collecting the texts for corpora creation, we built a crowd-sourcing portal called Wanca (Jauhiainen et al., 2019a; Jauhiainen et al., 2020).[17] The Wanca service enabled us together with a few collaborating language researchers and native speakers to easily inspect the web-pages tagged with minority languages by our language identifier. When identification mistakes were found, the wrongly set language labels were corrected manually using the service. In addition to helping us verify the language labels of the downloaded pages, Wanca functions as a collection of links for those interested in the Uralic minority languages. The service is currently maintained as a part of the Language Bank of Finland at the University of Helsinki.

The Wanca 2017 corpora are the product of a re-crawl performed by the SUKI project in October 2017. The target of the re-crawl was to download and check the availability of the then current version of the Wanca service of about 106,000 pages. This list of 106,000 http addresses was the result of several earlier web-crawls, in which we had identified the language of a total of 3,753,672,009 pages. We have listed the crawls with information about their target domains, date, and the number of pages processed in Table 1. In addition to our own crawls, we had identified the language of all the pages in the Common

[8] http://urn.fi/urn:nbn:fi:lb-2017021506

[9] https://www.statmt.org/europarl/

[10] https://wortschatz.uni-leipzig.de/en/download

[11] http://gtweb.uit.no/korp/

[12] http://gtweb.uit.no/f_korp/

[13] http://gtweb.uit.no/u_korp/

[14] http://urn.fi/urn:nbn:fi:lb-2019052402

[15] https://universaldependencies.org

[16] http://urn.fi/urn:nbn:fi:lb-2020021119

[17] http://wanca.fi/wanca/

Crawl archive[18] from December 2014 with almost two billion pages.

Name of crawl	Domains crawled	Date	Pages downloaded
SecondFinCrawl	.fi	22.5. – 9.6.2014	353,961,939
SweCrawl	.se	27.6. – 18.7.2014	308,130,342
NoCrawl	.no	2.8. – 23.8.2014	357,512,200
RuCrawl	.ru	7.8. – 4.9.2014	200,839,449
EeCrawl	.ee	10.9. – 14.9.2014	107,806,431
ThirdRuCrawl	.ru	17.9. – 23.9.2014	171,627,896
FourthRuCrawl	.ru	27.9. – 4.10.2014	115,419,359
FifthRuCrawl	.ru	4.10. – 22.10.2014	316,675,966
ThirdEeCrawl	.ee	15.10. – 22.10.2014	102,622,461
LVCrawl	.lv	29.10. – 18.11.2014	161,686,660
SecondNoCrawl	.no	29.10. – 20.11.2014	216,343,115
HuCrawl	.hu	29.10. – 26.11.2014	500,065,403
FinnishCrawl	.fi	18.11. – 26.11.2014	101,788,585
ComCrawl	.com, .ee, .fi, .hu, .lv, .no, .ru, .se	2.12.2014 – 20.1.2015	505,627,335
NLCrawl	.biz, .com, .org, .net	22.1. – 23.1.2015	233,564,868

Table 1: The web-crawls conducted by the SUKI-project prior to the 2017 re-crawl.

The re-crawl managed to download over 70% of the target urls. We processed the downloaded pages following the strategy presented by Jauhiainen et al. (2020) as follows.

First, all the text from each of the 78,685 downloaded pages was sent to a language set identification service we had set up for the task. The service used the HeLI language identification method together with the language set identification algorithm we had developed earlier (Jauhiainen et al., 2015c). The language set identification service used the latest language models of the SUKI project for a total of 399 languages or variants.[19] The code for the language identification service is available in GitHub with a GNU license.[20] We have not published the language models themselves as their purpose has always been to separate the small Uralic languages from the non-relevant languages and then discriminate between them. There are severe problems in discriminating between non-relevant languages, but sorting them out was not in the interest of the project as long as they did not interfere with the successful identification of the relevant languages. The training data for the ULI task will provide a better basis to train models for a language group-independent language set identifier destined for a more general use.

From the language set identified pages, we retained only those which had at least 2% text written in one of the minority Uralic languages. The relevant language that was most prominent was set as the identified language of the page. The retained pages contained a total of 1,515,068 lines and along with the lines, the identified language of the original page was kept. The lines were checked for duplicates, which left 446,233 unique lines. If the duplicates came from pages with different identified language, all those languages were set as the previously known languages of the line. Each line was then again sent to the language set identifier, which was only allowed to consider the previously known minority languages of the line as well as all non-relevant languages. Again only such lines were retained which included at least one relevant language, leaving 356,637 lines.

Next, a language-independent sentence extraction algorithm was run on each line. For this task, we had created a custom implementation of the sentence boundary disambiguation approach by Mikheev (2002) (Jauhiainen et al., 2019a). In total, 560,821 sentences were extracted using the algorithm with 477,109 of those being unique. After this, one more round of language set identifications was performed, this time for each unique sentence. Of the minority Uralic languages, the service was again allowed to consider only those in the list of the previously known languages of a sentence, but this time the absolute majority language of the identification was set as the language of the sentence. The resulting corpora contains 447,927 sentences in relevant languages divided as shown in the Wanca 2017 column of Table 2.

[18] https://commoncrawl.org

[19] The language models used in the project were semi-regularly updated using new or manually more checked corpora.

[20] https://github.com/tosaja/TunnistinPalveluMulti

	Wanca 2016	ULI 2020 training	Wanca 2017	ULI 2020 test
Finnic				
Estonian, Standard (**ekk**)	-	10,000	-	10,000
Finnish (**fin**)	-	1,000,000	-	10,000
Finnish, Kven (fkv)	2,156	2,156	1,499	23
Finnish, Tornedalen (fit)	5,203	5,203	4,517	100
Ingrian (izh)	81	81	80	-
Karelian (krl)	2,593	2,593	2,513	94
Liv (liv)	705	705	343	68
Livvi-Karelian (olo)	9,920	9,920	6,486	179
Ludian (lud)	771	771	411	185
Veps (vep)	13,461	13,461	9,122	2,453
Vod (vot)	20	20	11	-
Võro (vro)	66,878	66,878	61,430	443
Hungarian (**hun**)	-	1,000,000	-	10,000
Khanty (kca)	1,006	1,006	940	24
Mansi (mns)	904	904	825	1
Mari				
Mari, Hill (mrj)	30,793	30,793	22,986	18
Mari, Meadow (mhr)	110,216	110,216	38,278	3,768
Mordvin				
Erzya (myv)	28,986	28,986	16,273	1,153
Moksha (mdf)	21,571	21,571	15,170	724
Permian				
Komi-Permyak (koi)	8,162	8,162	6,104	-
Komi-Zyrian (kpv)	21,786	21,786	18,966	254
Udmurt (udm)	56,552	56,552	42,545	3,562
Saami				
Saami, Inari (smn)	15,469	15,469	14,405	228
Saami, Kemi (sjk)	19	19	-	-
Saami, Kildin (sjd)	132	132	59	13
Saami, Lule (smj)	10,605	10,605	5,644	400
Saami, North (sme)	214,226	214,226	165,009	6,009
Saami, Skolt (sms)	7,819	7,819	6,696	202
Saami, South (sma)	15,380	15,380	7,204	355
Saami, Ume (sju)	124	124	4	1
Samoyed				
Nenets (yrk)	443	443	407	58
Nganasan (nio)	62	62	-	-

Table 2: The number of sentences in Uralic languages for each dataset. The names of the relevant languages are in boldface.

4 The ULI 2020 shared task

The ULI 2020 shared task was organized as a part of the VarDial 2020 Evaluation Campaign. The evaluation campaign was the 7th incarnation of a series of shared tasks concentrating on close languages which have always incorporated some form of language identification tasks (Zampieri et al., 2014; Zampieri et al., 2015; Malmasi et al., 2016; Zampieri et al., 2017; Zampieri et al., 2018; Zampieri et al., 2019).

4.1 The dataset for the ULI shared task

The dataset for the ULI shared task consisted of two groups of languages: the relevant and the non-relevant. The relevant languages were the 29 minority Uralic languages listed in Table 2, which were present in the Wanca 2016 corpora (Jauhiainen et al., 2019a). The non-relevant languages were all the other languages for which at least two different datasets were downloadable from the Leipzig Corpora Collection (Goldhahn et al., 2012).

We used the whole Wanca 2016 corpora as training material for the for the relevant languages of the shared task and extracted a test set of new sentences from the Wanca 2017 corpora. The Wanca 2016 corpora is available from the Language Bank of Finland with a CC-BY license.[21] As Wanca 2017 was

[21] Helsingin yliopisto, FIN-CLARIN, Jauhiainen, H., Jauhiainen, T., & Lindén, K. (2019). Wanca 2016, source [text corpus].

not a real web-crawl, but only included downloading links already existing in the Wanca portal, it was in doubt how many completely new sentences the test set would have. For the ULI 2020 test set, we compared the Wanca 2017 corpora with the Wanca 2016 corpora and kept such sentences that were only found on the 2017 edition. This set including 25,547 sentences was then checked by us manually. We removed from the test set for relevant languages all obscure sentences as well as sentences that could have been incorrectly identified, concentrating on improving precision over recall. We were left with a total of 20,315 sentences divided between the minority Uralic languages as seen in the "ULI 2020 test" column of the Table 2.

In addition to the relevant languages, the training and test sets include sentences in 149 other languages from the Leipzig Corpora Collection. Following the goals of the SUKI-project, the three largest Uralic languages have been included in this category. The motivation for adding non-relevant languages to the shared task was to simulate the situation we faced when trying to find the texts written in minority Uralic languages on the Internet. For each page of text written in a relevant language we had found, we had identified the language of more than 50,000 pages. This kind of very unbalanced situation demands completely different levels of precision in identifying the relevant languages, when compared with any other language identification shared task so far (Grouin et al., 2011; Baldwin and Lui, 2010; Zubiaga et al., 2014; Solorio et al., 2014; Zampieri et al., 2014; Zampieri et al., 2015; Malmasi et al., 2016; Zampieri et al., 2017; Rangel et al., 2017; Ali et al., 2017; Zampieri et al., 2018; Zampieri et al., 2019). The download links for the training data for these non-relevant languages were distributed by the task organizers only to participating teams. In total, the training data for the task consisted of 63,772,445 sentences in non-relevant and 646,043 sentences in relevant languages, totaling 64,418,488 sentences. The list of the non-relevant languages is available at the Evaluation campaign website and the download links can be requested from the organizers. The Wanca 2017 corpora and the ULI test set will be published in the Language Bank of Finland with a CC-BY license after the shared task has been concluded.

4.2 Three tracks

The ULI 2020 shared task included three tracks. The tracks were not just about distinguishing between Uralic languages themselves, but also distinguishing the Uralic languages from the 149 non-relevant languages. The training and the test data for each of the tracks was the same and in each track every line in the test set was to be identified. The difference between the tracks was how the resulting scores were calculated, which significantly affects how the used classifying algorithms should be trained.

The first track of the shared task considered all the relevant languages equal in value and the aim was to maximize their average F_1-score. This is important when one is interested to find also the very rare languages included in the set of relevant languages. The results were calculated as macro-averaged F_1-scores over the small Uralic languages. In other words, for each of the 29 relevant languages present in the training set a separate recall and precision were calculated, even for those not present in the test set. The F_1-score for each language was calculated using Equation 1,

$$F_1(r,p) = \frac{2rp}{r+p} \qquad (1)$$

where r is the recall and p is the precision. If the correct number of true positives for a language was zero, then precision was 100% if no false positives were predicted. If false positives were predicted, the precision was zero. So, for those five languages (Ingrian, Vod, Komi-Permyak, Kemi Saami, and Nganasan) that were part of the training set, but did not appear in the test set, the recall was always 100% and precision was either 100% (if no instances of these languages were predicted in the test set) or 0% (if even one sentence was labeled as one of them). The result was the average of the F_1-scores of the 29 relevant languages. This means that, for example, predicting one false positive sentence for Ume Saami (which has only one sentence in the test set) is equal to predicting 6,009 false positives for North Saami (which has 6,009 sentences in the test set) as far as the results of the track were concerned.

Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2020022901.

The second track considered each sentence in the test set that is written in or is predicted to be in a relevant language as equals. The resulting F_1-score was calculated as a micro-F_1 over the sentences in the test set for sentences in the relevant languages as well as those that were predicted to be in relevant languages. When compared with the first track, this track gave less importance to the very rare relevant languages as their precision was not so important when the resulting F_1-score was calculated due to their smaller number of sentences. For example, predicting 6,009 false positives for North Saami in this track had 6,009 times the effect of predicting one false positive sentence for Ume Saami.

In the first two tracks, there was no difference between the non-relevant languages when the F_1-scores were calculated. The results were not affected, for example, if Norwegian sentences were identified as Danish or vice versa. The third track, however, did not concentrate on the 29 relevant languages, but instead the target was to maximize the average F_1-score over all the 178 languages present in the training set. The F_1-score was calculated as a macro-F_1 score over all the languages in the training set. This track was the language identification shared task with the largest number of languages to date (The ALTW 2010 shared task organized by Baldwin and Lui (2010) included 74 languages).

4.3 Baseline experiments

The baseline experiments were conducted using a language identifier based on the HeLI method (Jauhiainen et al., 2016). As a method, the HeLI method belongs to the generative classification methods and is a close relative to Naive Bayes. In the earlier VarDial shared tasks (Jauhiainen et al., 2015b; Jauhiainen et al., 2016; Jauhiainen et al., 2017a), we have successfully managed to compete almost at the same level as the best discriminative classification methods (Goutte and Léger, 2015; Malmasi and Dras, 2015; Çöltekin and Rama, 2016; Bestgen, 2017). In the HeLI method, each word in the mystery text has equal weight when determining the language of a text. Each word is divided into character n-grams, where the maximum length of the character sequences, n_{max}, is determined using the training and the development sets. Other tunable parameters include a cut-off, c, for the minimum frequency of features used as well as a penalty value, p, for unseen features. Instead of tuning the parameters using the ULI 2020 training set, we used the parameters we presented in Jauhiainen et al. (2017b): $n_{max} = 6$, $c = 0.0000005$, and $p = 7$. As we did in Jauhiainen et al. (2017b), we used the relative frequency of features as a cut-off instead of a raw frequency as the training corpora were of very different sizes. Only lowercased alphabetical characters were used in the language models. Due to HeLI using space character to separate words, there was a special 'sanity check' algorithm for texts including more than 50% CJK (Chinese-Japanese-Korean) characters, which gave all non-CJK languages a high penalty. The HeLI implementation used is almost exactly the same as the "TunnistinPalveluFast" available from GitHub.[22]

We did only one common run for all three tracks of the shared task. The results are listed in Table 3.

Track	F_1-score
ULI track 1 (ULI-RLE), relevant macro F_1	0.8004
ULI track 2 (ULI-RSS), relevant micro F_1	0.9632
ULI track 3 (ULI-178), macro F_1	0.9252

Table 3: The results of the baseline language identification experiments using the HeLI method.

Table 4 displays a confusion matrix showing two of the worst performing languages on Track 1: Ingrian and Votic. There were no real instances of Ingrian in the test set, but our baseline-identifier had identified three sentences of Ludian and one sentence of Karelian as Ingrian. These three languages are all closely related, but also one sentence of Sundanese was identified as Ingrian. The sentence in question is "Unggal lempir kawengku ku tilu padalisan." Both words "ku" and "tilu" are found in the Wanca 2016 corpus for Ingrian, which gives a hint of the reason for the mistake. Another language with an F_1-score of zero was Votic. Two Ludian sentences were identified as Votic, which is again understandable due to the languages being relatives, but also one sentence in Southern Sotho was identified as Votic: "Madinayne a ja dikokwanyana." As it happens, "a" is the most common word in the Wanca 2016 corpus for Votic and "ja" the sixth most common.

[22] https://github.com/tosaja/TunnistinPalveluFast

Table 4 also includes Tornedalen Finnish and Kven, two variants of Finnish used in Sweden and Norway, respectively. They are extremely close to the written versions of the Finnish dialects used in northern Finland. Relatively many Finnish sentences in the test set (fin_newscrawl_2017_10K-sentences) were identified as one of them. If the Finnish test set would have included social media texts instead of news articles, the confusion would have been much greater as standard written Finnish differs clearly from the written version of the colloquial Finnish.

Language	fin	**fit**	**fkv**	hat	**izh**	kpv	krl	lud	sot	sun	swe	**vot**
Finnish (fin)	*9,931*	53	7				3	1			1	
Tornedalen Finnish (fit)	5	*91*	2	1							1	
Kven (fkv)		3	*19*									
Haitian (hat)				*9,924*								
Ingrian (izh)												
Komi-Zyrian (kpv)		1				*246*						
Karelian (krl)	1				1		*80*					
Ludian (lud)	2				3			*144*				2
Southern Sotho (sot)									*9,962*			1
Sundanese (sun)				1	1				1	*5,451*		
Swedish (swe)	1										*9,981*	
Votic (vot)												

Table 4: Confusion matrix of some of the worst performing languages on Track 1.

Table 5 shows the languages which were confused with Võro, the worst performing language on Track 2. Võro is an extremely close language to Standard Estonian, both spoken in modern Estonia. None of the sentences in Võro were identified as Standard Estonian (ekk_web_2011_10K); however, over a thousand sentences (out of 10,000) in Standard Estonian (ekk_wikipedia_2016_10K) were identified as Võro. The only mistake in identifying sentences in Võro was when the sentence “”Õdaguhe” (film) ; 20:35 . ” was identified as Northern Azerbaijani. The number of false positives for Võro is explained by the domain difference between the Standard Estonian training and test sets. The training set for Standard Estonian was 10,000 sentences from news articles and the test set was 10,000 sentences from Estonian Wikipedia. The training set for Võro has a total of 66,878 sentences and 9,571 of those are from Võro language Wikipedia. The Võro and Standard Estonian Wikipedias discuss mostly the same named entities (foreign and domestic) and they were present only in the Võro training data, which resulted in a lot of sentences with those named entities being identified as Võro.

Language	azj	ekk	fin	gsw	hif	ita	lud	sun	tso	vec	**vro**	wuu
N. Azerbaijani (azj)	*9,896*											
Std. Estonian (ekk)	3	*8,717*	16	3	4	6	3			5	1,052	
Finnish (fin)			*9,931*				1				1	
Swiss German (gsw)				*9,409*	1	6	1			5	1	
Fiji Hindi (hif)			1	2	*9,246*	4			1		1	1
Italian (ita)				1	2	*8,723*			1	1,026	1	
Ludian (lud)			2	1			*144*				1	
Sundanese (sun)					2	1		*5,451*	4		1	
Tsonga (tso)									*9,991*		1	
Venetian (vec)				3		1,296				*747*	1	
Võro (vro)	1										*442*	
Wu Chinese (wuu)	1			2	9			8			1	*6,103*

Table 5: Confusion matrix for Võro, the worst performing language on Track 2.

To illustrate the identification errors in Track 3, we selected some of the worst performing languages and created a confusion matrix which is presented in Table 6. Bashkir and Tatar are closely related Turkic languages spoken in Russia. According to Tyers et al. (2012), their orthographical system are fairly different, which might indicate that the corpora used could be noisier than average. The extremely closely related languages Bosnian and Croatian have always been a problem for the non-discriminative HeLI method as is evidenced by the poor results for these languages in the DSL shared tasks of 2015, 2016, and 2017 (Jauhiainen, 2019). Wu Chinese was identified as Mandarin Chinese over 30% of the time. Character-based methods should be used instead of word-based methods when word-tokenization is a problem and the simple CJK algorithm included in the baseline-identifier just helps to correct some of the problems between CJK and non-CJK languages, but does not help in distinguishing between

CJK languages. The trio of close languages Indonesian, Javanese, and Sundanese got confused to the point of Indonesian being more often identified as Sundanese than Indonesian. Low German (nds-nl_wikipedia_2016_10K) was almost never identified as such (nds_wikipedia_2010_100K), but mostly as Limburgan (lim-nl_web_2015_300K). This seems to be due to the writing system of Low German being in flux and the nds.wikipedia[23] and nds-nl.wikipedia[24] being different entities.

Language	**bak**	**bos**	**cmn**	**hrv**	**ind**	**jav**	**lim**	**nds**	**sun**	**tat**	**wuu**
Bashkir (bak)	*6,961*									3,037	
Bosnian (bos)		*4,403*		5,593							
Mandarin Chinese (cmn)			*9,562*								273
Croatian (hrv)		1,134		*8,864*							
Indonesian (ind)					*3,102*	14			4,858		
Javanese (jav)					1,451	*4,626*			3,619		
Limburgan (lim)							*9,540*	10			
Low German (nds)							5,625	*182*			
Sundanese (sun)					87	4,330	1		*5,451*		
Tatar (tat)	3,784									*6,215*	
Wu Chinese (wuu)		1	3,610		1	2	1		8	1	*6,103*

Table 6: Confusion matrix of some of the worst performing languages by absolute numbers.

5 Conclusions and future work

In the beginning, we were worried about not getting enough new sentences from a simple re-crawl of the old addresses. In the end, the new sentences created an interesting setting for a language identification shared task. The three tracks highlighted different aspects of the problem of language identification.

The next edition of the ULI shared task will incorporate new sentences from the 2018 crawl performed by the SUKI project. Before processing the crawled material, we aim to improve our sentence extraction algorithm in such a way that it could allow sentences to span line-breaks. Also, as pointed out by one of the reviewers, it would be a good idea to manually inspect at least a random subset of the Wanca 2017 corpora in order to objectively assess the reliability of the language annotation process. Unlike the 2017 re-crawl, the 2018 crawl was a real crawl going beyond the addresses stored in the Wance service. Thus, we expect to find more new sentences after we process the material using the improved process.

Acknowledgements

The work presented here was conducted with funding by the University of Helsinki and the Academy of Finland. The SUKI project was funded by the Kone Foundation from 2013 to 2019. We thank the anonymous reviewers for their questions and suggestions.

References

Daniel Abondolo and Riitta-Liisa Valijärvi, editors. forthcoming. *The Uralic Languages*. London: Routledge, 2nd edition.

Daniel Abondolo, editor. 1998. *The Uralic languages*. London: Routledge.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322.

Timofey Arkhangelskiy. 2019. Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140. Tartu.

Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors. forthcoming. *The Oxford Guide to the Uralic Languages*. Oxford: Oxford University Press.

[23] https://nds.wikipedia.org/wiki/Plattdüütsch

[24] https://nds-nl.wikipedia.org/wiki/Nedersaksisch

Timothy Baldwin and Marco Lui. 2010. Multilingual Language Identification: ALTW 2010 Shared Task Dataset. pages 5–7, Melbourne, Australia.

Yves Bestgen. 2017. Improving the Character Ngram Model for the DSL Task with BM25 Weighting and Less Frequently Used Feature Sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain.

Cagri Çöltekin and Taraka Rama. 2016. Discriminating Similar Languages: Experiments with Linear SVMs and Neural Networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 15–24, Osaka, Japan.

Timothy Feist. 2015. *A grammar of Skolt Saami*. Number 273 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Cyril Goutte and Serge Léger. 2015. Experiments in Discriminating Similar Languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 78–84, Hissar, Bulgaria.

Lenore A Grenoble. 2003. *Language policy in the Soviet Union*. Kluwer Academic Publishers.

Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2011. Présentation et Résultats du Défi Fouille de Texte DEFT2010 Où et Quand un Article de Presse a-t-il Été Écrit? In *Actes du sixième Défi Fouille de Textes*, pages 3–14, Montpellier, France.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2015a. The Finno-Ugric Languages and The Internet Project. In *Proceedings of the 1st International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2015)*, number 2 in Septentrio Conference Series, pages 87–98.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2015b. Discriminating Similar Languages with Token-Based Backoff. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 44–51, Hissar, Bulgaria.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2015c. Language Set Identification in Noisy Synthetic Multilingual Documents. In A. Gelbukh, editor, *Proceedings of the Computational Linguistics and Intelligent Text Processing 16th International Conference, (CICLing 2015)*, Part I of Lecture Notes in Computer Science, pages 633–643, Cairo, Egypt. Springer.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. HeLI, a Word-Based Backoff Method for Language Identification. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 153–162, Osaka, Japan.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017a. Evaluating HeLI with Non-Linear Mappings. In *Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, pages 102–108, Valencia, Spain.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017b. Evaluation of Language Identification Methods Using 285 Languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa 2017)*, pages 183–191, Gothenburg, Sweden.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Linden. 2019a. Wanca in Korp: Text corpora for underresourced Uralic languages. In Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen, and Katja Västi, editors, *Proceedings of the Research data and humanities (RDHUM) 2019 conference*, number 17 in Studia Humaniora Ouluensia, pages 21–40, Finland. University of Oulu.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019b. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2020. Building web corpora for minority languages. In *Proceedings of the 12th Web as Corpus Workshop*, pages 23–32.

Tommi Jauhiainen. 2019. *Language identification in texts*. Ph.D. thesis, University of Helsinki, Finland.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. pages 79–86, Phuket, Thailand.

Irina Krylova, Boris Orekhov, Ekaterina Stepanova, and Lyudmila Zaydelman. 2015. Languages of Russia. In *Russian Summer School in Information Retrieval*, pages 179–185.

Shervin Malmasi and Mark Dras. 2015. Language Identification using Classifier Ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 35–43, Hissar, Bulgaria.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Osaka, Japan.

Andrei Mikheev. 2002. Periods, Capitalized Words, etc. *Computational Linguistics*, 28(13):289–318.

Gordon Mohr, Michael Stack, Igor Rnitovic, Dan Avery, and Michele Kimpton. 2004. Introduction to Heritrix. 4th International Web Archiving Workshop (at ECDL2004).

Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*, volume 65. Walter de Gruyter GmbH & Co KG.

Boris Orekhov, Irina Krylova, I. Popov, L. Stepanova, and Lyudmila Zaydelman. 2016. Russian minority languages on the web. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016*, pages 498–508.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian universal dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Tommi A Pirinen. 2019. 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)*, pages 132–136.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September. CEUR-WS.org.

Jack Michael Rueter and Francis M Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. In *International Workshop for Computational Linguistics of Uralic Languages*.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020. On the questions in developing computational infrastructure for Komi-Permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 15–25.

Jack Rueter, Mika Hämäläinen, and Niko Partanen. in press. Open-source morphology for endangered mordvinic languages. In *Proceedings of the 2nd Workshop for Natural Language Processing Open Source Software (NLP-OSS)*.

Florian Siegl and Michael Rießler. 2015. Uneven steps to literacy. In *Cultural and Linguistic Minorities in the Russian Federation and the European Union*, pages 189–230. Springer.

Denis Sinor, editor. 1988. *The Uralic languages: Description, history and foreign influences*, volume 1. Brill Academic Publishers.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October.

Francis Tyers and Mariya Sheyanova. 2017. Annotation schemes in north sàmi dependency parsing. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 66–75.

Francis M Tyers, Jonathan North Washington, Ilnar Salimzyanov, and Rustam Batalov. 2012. A prototype machine translation system for tatar and bashkir based on free/open-source components. In *First Workshop on Language Resources and Technologies for Turkic Languages*, page 11.

Beáta Wagner-Nagy. 2018. *A grammar of Nganasan*. Brill.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A Report on the DSL Shared Task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL Shared Task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 1–9, Hissar, Bulgaria.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešic, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–15, Valencia, Spain.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. 2014. Overview of TweetLID: Tweet Language Identification at SEPLN 2014. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 1–11, Girona, Spain, September.

Ènn Èrnits. 2006. Ob oboznačenii zvukov v vodskom literaturnom âzyke. *Linguistica Uralica*, 42(1):1–9.