Improving Language Coverage on HeLI-OTS

Jauhiainen, Tommi

2024

Jauhiainen, T & Linden, K 2024, Improving Language Coverage on HeLI-OTS. in M Melero, S Sakti & C Soria (eds), Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024. International conference on computational linguistics, LREC proceedings, European Languages Resources Association (ELRA) , Paris, pp. 115-125, Annual Meeting of the ELRA-ISCA Special Interest Group on Under-resourced Languages (Sigul 2024), Torino, Italy, 20/05/2024.

http://hdl.handle.net/10138/576188

Improving Language Coverage on HeLI-OTS

Tommi Jauhiainen and Krister Lindén
Department of Digital Humanities
University of Helsinki
{firstname.lastname}@helsinki.fi

Abstract

In this paper, we add under-resourced languages into the language repertoire of an existing off-the-shelf language identifier, HeLI-OTS. Adding more languages to a language identifier often comes with the drawback of lessened accuracy for the languages already part of the repertoire. We aim to minimize this effect. As sources for training and development data in the new languages, we use the OpenLID and FLORES-200 datasets. They are openly available high-quality datasets that are especially well-suited for language identifier development. By carefully inspecting the effect of each added language and the quality of their training and development data, we managed to add support for 20 new under-resourced languages to HeLI-OTS without affecting the performance of any existing languages to a noticeable extent.

Keywords: language identification, text corpora

1. Introduction

Language identification (LI) involves figuring out the language in which a document or a portion of it is written. The techniques for automatically determining the language of digital texts have been developed for over five decades. Over time, the importance of language identification as a crucial preliminary step has increased, especially as natural language processing (NLP) technologies have become integral to everyday applications (Jauhiainen et al., 2019b; Jauhiainen, 2019; Jauhiainen et al., 2024). For instance, to carry out machine translation of text, it is necessary to know the source language. Without an automated system for identifying languages, users must manually specify the text's language. Google Translate is an example of a platform that has integrated language identification capabilities.

This paper details handling the workflow of adding languages to the HeLI-OTS off-the-shelf language identifier (Jauhiainen et al., 2022a). Section 2 introduces HeLI-OTS and mentions other off-the-shelf language identification tools. Section 3 details the OpenLID and FLORES-200 corpora we use to improve the language coverage on HeLI-OTS. We introduce the workflow of adding languages to HeLI-OTS in Section 4, and in Sections 5 and 6, we introduce the added languages and their statistics as well as give some observations we made while adding them to the HeLI-OTS repertoire. In Section 7, we evaluate HeLI-OTS with the added languages on the FLORES-200 test partition and compare its results with the state of the art. In the last Section, we discuss the findings and draw conclusions.

2. Previous Work

HeLI-OTS is based on the HeLI language identification method we have been developing for more than a decade (Jauhiainen et al., 2016). The HeLI method has proven to be robust in handling difficult situations with, e.g., a large number of languages and out-of-domain target texts (Jauhiainen et al., 2017).

The first version of the HeLI-OTS off-the-shelf language identifier was published in Zenodo in May 2021.[1] Since then, we have been improving the quality of existing language models and adding new functionality to the software which is currently on its fifth version, 1.5, published in November 2023 (Jauhiainen and Jauhiainen, 2023). The 200 language repertoire was carefully curated for the first version (Jauhiainen et al., 2022a). The repertoire has remained identical since the first version, even though we have improved and added new training and development material for the existing languages. The development of the language identifier has been conducted, e.g., as part of improving the resource publishing pipeline of the Language Bank of Finland (Jauhiainen et al., 2022b; Dieckmann et al., 2023) or participating in language identification shared tasks (Jauhiainen et al., 2023).[2] For version 1.5, we added a language set identification functionality using a method we had developed earlier (Jauhiainen et al., 2015).

This paper details the first occasion of expanding the language repertoire beyond 200 languages.

The first widespread off-the-shelf language identification tool was TextCat (van Noord, 1997) us-

[1] https://zenodo.org/doi/10.5281/zenodo.4780897

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

SIGUL2024 Workshop, pages 115–125

ing the method developed by Cavnar and Trenkle (1994) with 76 languages. The next widely used tool that replaced TextCat was langid.py, which had models for 97 languages (Lui and Baldwin, 2012). Currently, the most widely used tools are based on the fastText method (Joulin et al., 2017). The first fastText-based language identifier was published in 2018, including models for 178 languages.[3] The second version of the Facebook/Meta AI Research published language identifier was unveiled as part of their No Language Left Behind (NLLB) initiative in 2022 (NLLB Team, 2022). It has language models for 218 languages.[4] In 2023, Burchell et al. (2023) published another fastText based language identifier for 201 languages[5] and evaluated its accuracy against the NLLB version.

3. Source Corpora

Good quality monolingual language data is surprisingly difficult to acquire in large amounts. Kreutzer et al. (2022) evaluated five multilingual corpora and found severe quality-related issues, especially with under-resourced languages.

Heeding the advice from the lessons learned by Kreutzer et al. (2022), Burchell et al. (2023) decided to avoid web-crawled datasets when creating a new dataset for language identification purposes. When they published their OpenLID language identifier and the accompanying dataset for 201 languages, we decided that we should try to use the dataset to enlarge the language repertoire of our off-the-shelf language identifier. Burchell et al. (2023) chose the 201 languages so that they were the same as in the FLORES-200 dataset[6] (Guzmán et al., 2019; Goyal et al., 2021; NLLB Team, 2022) so that they could use it for verifying and evaluating the resulting classifier. The OpenLID dataset contains 121 million lines of text spanning from 532 lines for South Azerbaijani to 7.5 million lines for English. The majority of the texts in the dataset originate from news sites, Wikipedia, or religious texts (Burchell et al., 2023).

The FLORES-200 dataset has two parts: one for development "dev" and one for testing "devtest". Both contain material for 196 languages, eight of which have two versions with differing scripts. Each of the 204 language-script combinations has 997 lines for development and 1012 lines for testing per language.

[3] https://fasttext.cc/docs/en/language-identification.html

[4] https://github.com/facebookresearch/fairseq/tree/nllb

[5] https://github.com/laurieburchell/open-lid-dataset

[6] https://github.com/facebookresearch/flores/tree/main/flores200

4. Adding Languages

We wanted to begin adding languages so that the training data would be of the highest quality. In order to attain this, we inspected which languages had scored the best in the evaluation carried out by Burchell et al. (2023).[7] As the evaluation measure, they used the F1 score (or F-score) which is a measure widely used in the evaluation of language identification performance (Jauhiainen et al., 2024; Aepli et al., 2023). F-score combines both recall and precision. For each language, recall indicates the percentage of how many of the lines in the language are identified as such. The lines identified as some other languages or as no language at all count as false negatives. Precision tells which percentage of the lines identified as the language are actually in that language. The lines in other languages are then called false positives. A perfect F-score can be attained only when both recall and precision are perfect.

For the first batch (Section 5) of added languages, we considered all those twelve languages that had attained a perfect F-score and were not yet part of the HeLI-OTS language repertoire: Tosk Albanian, Central Aymara, Bashkir, Central Kurdish, Jingpho, Halh Mongolian, Odia, Plateau Malagasy, Ayacucho Quechua, Santali, Shan, and Waray. When looking at these languages, we noticed that we already had the macrolanguage listed for Tosk Albanian, Halh Mongolian, Odia, Plateau Malagasy, and Ayacucho Quechua. The Open-LID language repertoire did not include any other languages belonging to the respective macrolanguages, so we could not add them as a macrolanguage and an individual language belonging to it cannot reside on the same level in the identification hierarchy. We were left with seven new languages. We began processing them into the repertoire, starting from the ones with the most speakers according to sources linked to by the ISO 639-3 standard website,[8] mainly Wikipedia.

For the second batch (Section 6), we chose to inspect the 22 languages which had attained F-scores higher or equal to 0.998: Achinese, North Azerbaijani, Southwestern Dinka, Fon, Friulian, West Central Oromo, Northern Kurdish, Central Kanuri, Ligurian, Latgalian, Standard Latvian, Dholuo, Nepali, Nuer, Pangasinan, Southern Pashto, Samoan, Serbian, Tigrinya, Twi, Eastern Yiddish, and Yoruba. North Azerbaijani, Nepali, Latgalian, Standard Latvian, Serbian, and Eastern Yiddish were part of a macrolanguage that was already part of the HeLI-OTS language repertoire. For the remaining 16 languages, we again checked

[7] https://github.com/laurieburchell/open-lid-dataset/blob/main/languages.md

[8] https://iso639-3.sil.org

the number of their speakers and began processing them from the highest to the lowest. We continued until we reached 20 new languages. Friulian, Ligurian, and Samoan were left to be added in the future.

Adding a language to HeLI-OTS begins by using the then-current version to identify the language of each line of the training and development data for the candidate language and then manually inspecting the results. Severe foreign language incursions typically have a high confidence score, which is why we usually filter out lines with high confidence scores at this stage. Then, we add the development data to the HeLI-OTS internal test set and create language models for the candidate language. At the beginning of the process, the internal test set had 1,239,621 lines of text for the 200 languages. Then, we evaluate the internal test set using HeLI-OTS with the additional language and compare the results with those of the previous internal evaluation. Then, the internal test set is used to generate confidence thresholds for HeLI-OTS so that unnecessary false positives are avoided. Currently, the confidence thresholds for each language are the lowest confidence scores with which part of the corresponding language's test data has been correctly identified. In HeLI-OTS, the confidence score is the difference between the internal scores of the best and second-best guessed language (Jauhiainen et al., 2019a). HeLI-OTS can tag a text as written in an undetermined language "und" in two situations. The first is when the text does not contain any characters belonging to the character set of any language but consists only of characters such as numbers or punctuation. The second case is when confidence thresholds are used, and the confidence score for the text is lower than the threshold set for the most probable language.

Table 1 shows statistics for each of the 20 new languages added to HeLI-OTS as part of the work described in this paper. The first column gives the ISO 639-3 code for each of the languages, and the languages are listed in the same order as they appear in the two following sections. The second column indicates the number of lines available for the language as training data in the OpenLID corpus, and the next column tells how many of those lines we actually used as training data for the corresponding language in the HeLI-OTS. Each of the languages has 997 lines of development data in the FLORES-200 dataset. The "Retained Testing Size" column tells how many of those lines we added to the internal test set. The second to last column gives the F-score for each language on the internal test set without the use of confidence thresholds. These results are generated when we are determining the confidence thresholds. The last column gives the F-score with the confidence scores for each language. In this table, both scores are from the point of time when the corresponding language (and all the languages appearing before it on the list) had been just added to the HeLI-OTS language repertoire.

5. First Batch

Santali [sat] Santali language belongs to the Austro-Asiatic languages and is spoken in India, Bangladesh, and Nepal and is categorized as "Institutional" in language vitality by Ethnologue (Eberhard et al., 2023).[9] It is spoken by more than 6 million people (Akhtar et al., 2017). The Santali corpus in the OpenLID dataset included 8,875 lines, of which the language was left undetermined by HeLI-OTS 8,773 times. The Santali uses a new writing system as far as HeLI-OTS is concerned, and thus, most of the lines have not been mapped to any languages. The lines identified as something else contained some text, mostly in Latin characters. However, there were nine lines identified as Oriya, which is written using a completely different writing system that could visually be confused with the one used by Santali. For our training material, we decided to keep only those lines that were left undetermined by HeLI-OTS. For our internal test set, we kept all the 997 lines even though some of them contained Latin characters in addition to the characters of the new writing system.

Central Kurdish [ckb] The Central Kurdish language is one of the individual languages belonging to the Kurdish macrolanguage. It is one of the official national languages of Iraq (Eberhard et al., 2023).[10] The language, also known as Sorani, was spoken by c. 7 million people in 2015 (Hassani et al., 2016). Of the 17,792 lines of Central Kurdish (written using the Arabic script) in the OpenLID dataset, 12,045 were identified as Iranian Persian, 5,025 were left undetermined, and the rest were tagged with an assortment of languages, including 37 lines identified as written in Arabic. After manual inspection, it seemed that at least the Arabic-identified lines actually contained text written in Arabic. They were mostly titles of books and lists of their authors. We decided to keep all the lines left undetermined, and those Iranian Persian lines with confidence score less than 1.0. The lines with a low confidence score are less likely to actually be written using the language indicated. We used the same indicators when selecting lines from the FLORES 200 development set into our internal test set.

[9] https://www.ethnologue.com/language/sat/

[10] https://www.ethnologue.com/language/ckb/

ISO 639-3	OpenLID training size	Retained training size	Retained testing size	F-score without confidence	F-score with confidence
sat	8,875	8,773	997	1.0	1.0
ckb	17,792	16,393	905	0.9994	1.0
shn	21,051	18,868	736	1.0	1.0
war	282,772	250,148	949	0.9953	0.9958
ayr	142,628	110,908	837	1.0	1.0
bak	65,942	49,755	924	0.9908	0.9919
kac	11,365	11,364	997	0.9995	1.0
yor	531,904	526,661	997	0.9990	0.9990
gaz	335,769	330,651	997	1.0	1.0
kmr	15,490	13,779	997	0.9911	0.9925
pbt	63,256	62,229	775	0.9955	0.9994
twi	545,217	540,367	980	0.9990	0.9990
knc	6,256	5,933	963	1.0	1.0
tir	333,639	331,176	997	0.9990	0.9995
dik	25,911	25,783	985	1.0	1.0
luo	138,159	137,579	994	0.9980	1.0
ace	18,032	16,692	992	1.0	1.0
fon	31,875	31,048	997	0.9985	0.9990
pag	294,618	289,594	934	0.9952	0.9979
nus	6,295	4,330	996	0.9995	1.0

Table 1: Language addition to HeLI-OTS: corpus sizes and language-specific F-scores.

Shan [shn] Shan language is mostly spoken in Myanmar and by less than 5 million people worldwide (Eberhard et al., 2023).[11] It is written using the same orthography as Burmese, but the two languages are unrelated. So far, Burmese has been the only language using these Unicode characters, which led the Shan texts from both the OpenLID and FLORES-200 corpora to be mostly identified as Burmese using the HeLI-OTS. Out of the 21,051 lines of Shan in the OpenLID, 18,868 lines were identified as Burmese, 2,122 were left undetermined, and the rest, c. 60, were tagged with 10 Latin character-based languages. The latter group contained lines consisting only or mostly of text with Latin characters, and the lines in the undetermined category contained several words written in Latin characters as well. After inspecting the results, we decided to use only the lines identified as Burmese in our training corpus for Shan. Similar phenomena prevailed in the development part of the FLORES 200 dataset, except that additionally, most of the lines identified as Burmese contained at least one word written using Latin characters. However, we still incorporated all the lines tagged with Burmese into our internal test set. After the addition, both Burmese and Shan were 100% correctly identified, even without using confidence thresholds.

Waray (Philippines) [war] The Malayo-Polynesian Waray or Waray-Waray language is spoken by less than 3 million people, mostly residing in the Philippines (Eberhard et al., 2023).[12] The OpenLID corpus has 282,772 lines of text for Waray. HeLI-OTS identified 196,367 of those lines as Cebuano, 27,397 as Tagalog, and 9,381 as Central Bikol. 44,644 lines were left undetermined, and the remaining 4,983 lines were divided between 104 other languages. The 997 Waray texts from the development partition of FLORES-200 were identified as Cebuano 776 times, as Tagalog 72 times, and as Central Bicol only three times. 143 lines were left undetermined, and three lines were identified as two other languages. From both datasets, we decided to retain those lines identified with less than a 1.0 confidence score as Cebuano or Tagalog, as well as the lines left undetermined. When calculating the confidence scores, Waray reached an F-score of 0.9953 on the internal test set, which was above the average of 0.9928 for all 204 languages. It had two false negatives and seven false positives. Using the confidence threshold took away one of the false positives.

Central Aymara [ayr] Central Aymara belongs to the Aymara macrolanguage. It is spoken by less than 1.5 million speakers in total, two-thirds of whom reside in Bolivia (Eberhard et al., 2023).[13] Aymaran languages do not have any close relatives in the HeLI-OTS language repertoire. The Aymaran training corpus was tagged to be written in 118 dif-

[11] https://www.ethnologue.com/language/shn/

[12] https://www.ethnologue.com/language/war/

[13] https://www.ethnologue.com/language/aym/

ferent languages in addition to being tagged as undetermined. Of the 142,628 Aymaran lines, 74,953 were left tagged as undetermined and 26,096 as Quetchuan, which is a language spoken partly in the same geographical area. The next most tagged languages were Swahili (5,404) and Waray (5,364), which neither originate from the same continent. Most of the lines tagged with these four identifiers seemed to contain well-formed sentences, even though some of them seemed to contain much bible-related vocabulary. The fifth most common language was Spanish, with 3,604 lines, most of which actually contained Spanish words, and some were completely written in Spanish. This was expected for a language from this area. Previously, we have spent much effort cleaning Spanish out of the HeLI-OTS Guarani training data (Jauhiainen et al., 2023). As training material for HeLI-OTS, we kept the lines tagged as Quetchua, Swahili, and Waray with confidence scores less than 1.0 in addition to all the lines tagged as undetermined. For our internal test set, we took the lines from the FLORES 200 development set, which were tagged as undetermined or as Quechua (with less than a 1.0 confidence score).

Bashkir [bak] Bashkir, with around 1.2 million speakers, belongs to the Uralian subgroup of the Western Turkish language family (Eberhard et al., 2023).[14] Among the four languages belonging to this subgroup is Tatar, which is already part of the HeLI-OTS language repertoire. Of the 65,942 lines of Bashkir in the OpenLID dataset, 52,856 were identified as Tatar, 6,023 as Kazakh, 4,492 were left undetermined, and the rest were divided between 44 different languages. The Kazakh-identified lines seemed to be mostly very short, self-repeating descriptions of places. Also, the lines tagged as undetermined seemed to be very short template-like texts. The development set from FLORES-200 contained 997 lines tagged as Bashkir, of which 964 were identified as Tatar. From both datasets, we decided to keep only the lines that had been identified as Tatar. As Tatar is such a close relative to Bashkir, we decided to take those Tatar-identified lines that had a confidence score of less than 2.0 instead of the 1.0 we used in similar situations previously. Without using confidence thresholds, four of the Bashkir test lines were identified as something else than Bashkir, and Bashkir had attracted 13 false positives. The F-score for Bashkir was 0.9908, and the Tatar F-score dropped from 0.9996 to 0.9989. We deemed this a low enough price to pay, considering that there is now a new pair of close relatives within the language repertoire. With confidence thresholds, the F-scores were 0.9919 for Bashkir and 0.9990 for Tatar.

Jingpho [kac] Jingpho language belongs to the Tibeto Burman group and has no close relatives in the current HeLI-OTS language repertoire. It is written using the Latin alphabet and is spoken by less than 1 million speakers, mostly residing in Myanmar. Quickly browsing through lines in the training data after preliminary language identification, it seemed that there were few foreign language incursions in the text except the one line identified as English, which consisted mostly of English words. The same seemed to be true for the test data. We left out only the English-identified sentence and kept the rest of the lines for both data sets. Without confidence thresholds, Jingpho attracted one false positive identification, and even that was handled with thresholds.

6. Second Batch

Yoruba [yor] The 531,904 lines of the Yoruba training corpus were initially tagged with 125 different language codes, mostly with “und” for undetermined. The next most numerous tag was that of Irish, a completely unrelated language that was not really present at all. Inspecting the top languages, only English seemed to be actually present in large numbers. We decided to leave out all the 1,348 lines identified as English. Also, some of the 157 lines identified as Spanish were completely written in Spanish, so we left them out as well. Of the other than English and Spanish lines, we kept those with confidence scores less than 1.0. All of the 997 lines of the development set seemed to be okay; even the one line identified as Spanish did not seem to contain any foreign parts. We kept all the development lines for internal testing. Without confidence thresholds, Yoruba got one false negative and one false positive identification with an F-score of 0.9990, which also remained while using the thresholds.

West Central Oromo [gaz] Out of the 335,769 lines for training, 201,198 were tagged as undetermined. 43,328 lines were identified as Somali. According to Glottolog, both languages belong to the Mainstream Lowland East Cushitic group, along with 19 other languages.[15] Oromo is also spoken in the area of modern-day Somalia, so it is possible that the collection could contain some text in Somali. The next most common language was Finnish, which is a completely unrelated language, and we did not see any sign of it on the lines identified as such. Then, we proceeded to check for

[14] https://www.ethnologue.com/language/bak/

[15] https://glottolog.org/resource/languoid/id/main1283

languages that we have many times witnessed as incursions in other languages. The 1,411 lines identified as Italian seemed to be mostly short ones containing two or three words inside the parenthesis, so we decided to leave them out. Some of the 623 lines identified as English were completely written in English, so we left them out as well. After perusing the lines identified as Somali, we once again decided to keep those lines with confidence scores lower than 1.0 from the other than English- and Spanish-identified lines. The development set seemed to be of high quality, and we kept all the lines.

Northern Kurdish [kmr] Northern Kurdish belongs to the Kurdish macrolanguage, which belongs to the Northwestern Iranian language group of the Indo-European language family.[16] HeLI-OTS previously contains the Southern Zazaki language from this language group, which is also written similarly using Latin characters as the Nothern Kurdish data in the OpenLID data set. Of the 15,490 lines in the training set, 12,279 were identified as Southern Zazaki and 1,539 lines were left undetermined. Furthermore, 804 lines were identified as Turkish, which is a language used in close geographical proximity. Apart from the 17 lines identified as English, the text seemed to be of good quality. We retained all the lines left undetermined and all non-English identified lines with confidence scores less than 1.0. With a similar distribution for identified languages, the development set seemed of good quality, so we kept it all. Without confidence thresholds, Northern Kurdish attracted 18 false positives. This was a more significant number than we had seen so far in these experiments, so we decided to take a look at the results. 15 of the 18 lines were tagged with Southern Zazaki and looked rather well formed. The F-score for Southern Zazaki dropped from 0.9985 to 0.9966, so it was still very acceptable. Using confidence thresholds took away three false positives from Northern Kurdish.

Southern Pashto [pbt] Southern Pashto belongs to the Pushto macrolanguage. It belongs to the Eastern Iranian subgroup of Indo-European languages.[17] HeLI-OTS already contains the Ossetic language, which belongs to the same group. However, our Ossetian training data is written in Cyrillic as opposed to the Arabic script used for Southern Pashto in the OpenLID dataset. The 63,256 lines were identified as Iranian Persian 44,094 and left undetermined 16,171 times. Iranian Persian belongs to the Western Iranian language group and is rather closely related and written using the same writing system. We decided to keep all lines with identification confidence of less than 1.0. The development data included many lines with Latin characters, which we decided to filter out. Southern Pashto got seven false positives without confidence thresholds, and with the thresholds, only one false positive remained.

Twi [twi] Twi belongs to the Akan macrolanguage and to the Atlantic-Congo language family without any close relatives in the HeLI-OTS language repertoire. In the development set, Twi was most often identified as Dimli, which is a completely unrelated Indo-European language. In the development set, some lines were identified as English or Italian due to either actual incursions or a list of names. We decided to filter these languages out of the dataset. Also, the training data has some lines that included a great deal of English, which were filtered out. For the internal test set, Twi got two false positives with and without confidence thresholds.

Central Kanuri [knc] Central Kanuri belongs to the Kanuri macrolanguage belonging to the Nilo-Saharan language family.[18] It does not have any close languages in the HeLI-OTS language repertoire. The 6,256 lines of texts were left undetermined 3,701 times and then identified as Twi 404 and Dimli 394 times, languages which belong to two completely other language families. The 313 lines identified as English contained pieces of English sentences. We filtered out the English sentences and kept all other lines with confidence lower than 1.0. We filtered the English-identified lines out of the development set as well.

Tigrinya [tir] Tigrinya is an Afro-Asiatic language written in the same script as Amharic, which is already present in the HeLI-OTS language repertoire. Of the 333,639 lines in the OpenLID dataset, 331,176 were identified as Amharic. As there were no competitors in the repertoire, Amharic received very high confidence scores for all Tigrinyan sentences. All the lines identified as something else contained Latin characters in addition to the Ethiopian script or did not contain text written in the correct script at all. All the 997 lines of the development set were identified as Amharic. From both files, we kept only the lines identified as Amharic. Without confidence thresholds, Tigrinya attracted two false positives from Amharic, which dropped from a perfect F-score to 0.9999. One of the two false positives was taken away when thresholds were used.

[16] https://www.ethnologue.com/subgroup/21/

[17] https://www.ethnologue.com/subgroup/18/

[18] https://www.ethnologue.com/subgroup/767/

Southwestern Dinka [dik] Southwestern Dinka is part of the Dinka macrolanguage belonging to the Eastern Sudanic group of the Nilo-Saharan language family.[19] It does not have any close relatives among the HeLI-OTS language repertoire. Of the 25,911 lines of data in OpenLID, 17,706 were left undetermined, and 2,270 were identified as Dimli from the Indo-European language family. The lines identified as the top languages seemed good, but lines tagged as English again sometimes contained snippets of the foreign language. The same was true with the development set from FLORES-200. We filtered out English-identified lines from both sets and kept all undetermined lines and other lines with confidence scores less than 1.0. For the test set, we kept all lines except the 12 English-identified lines.

Dholuo, Luo (Kenya and Tanzania) [luo] Luo is also from the Eastern Sudanic group of the Nilo-Saharan language family. Of the 138,159 lines of data in the OpenLID dataset, 64,808 were left undetermined, 10,341 were identified as Dimli, and 8,772 were identified as Esperanto. The rest of the lines were divided between 47 other languages. The development lines were identified as a similar collection of seemingly random languages starting from Tagalog after undetermined lines. Lines identified as English in the training set once more included some completely English sentences. The three lines identified as English on the test set contained some English words. We filtered out the English lines and kept the rest, again filtering out those with confidence higher or equal to 1.0 in the training set. Without confidence thresholds, Luo got four false positive identifications, but after introducing the thresholds, it received a perfect F-score.

Achinese [ace] Achinese belongs to the Malayo-Chamic language group within the Austronesian language family. HeLI-OTS currently includes the Malaysian macrolanguage in its repertoire, and it can be considered a language that is close to Achinese. Of the 18,032 lines of the OpenLID dataset, 13,016 were left undetermined, and 1,181 were identified as Malaysian macrolanguage. On the development data from FLORES-200, the Malaysian macrolanguage did not make the top 10 languages, with only five lines out of 997. The other higher-ranked languages were much more similarly situated in the rankings. The Malaysian identified lines were also rather confident, unlike with the other language labels, and could be ranked out by using the 1.0 confidence filter as with previously processed languages. Again, the 105 English-identified lines contained a great deal of English, which we filtered out completely. There were no English-identified lines on the test set. This time, we also used the confidence threshold of 1.0 when filtering the test lines.

Fon [fon] Fon is a language belonging to the Volta-Congo group of the Niger-Congo language family. Both Yoruba and Twi, which we added earlier, belong to the same language group. The 31,875 lines in the training data were left undetermined 18,625 times. They were identified as Yoruba 9,615 times and as Twi 1,696 times. The 87 lines identified as French and the seven lines identified as English contained clear passages written in the respective languages. We filtered out English- and French-identified lines and lines with confidence scores of 1.0 or higher from the training data. The test data seemed of better quality; it was all retained. Fon got two false negatives and one false positive without the confidence thresholds. Using the threshold took the false positive identification away.

Pangasinan [pag] Pangasinan belongs to the Malayo-Polynesian language group of the Austronesian language family. From that group, HeLI-OTS already includes several languages, e.g., Tagalog, Kapampangan, Cebuano, and Central Bikol. We followed the previous examples and noticed that the English-identified lines were mostly English. We also decided to leave out lines identified as Spanish and French as well as all other lines with confidence equal to or higher than 1.0. Also, the English-identified lines in the development set included heavy code-switching, and we decided to leave them out of the test set. Without confidence thresholds, Pangasinan reached an F-score of 0.9952 with two false negatives and seven false positives. This must be considered a very good result, considering the nature of heavy code-switching in languages used in the Philippines. Using confidence thresholds, the F-score rose to 0.9979 with only two false positive identifications.

Nuer [nus] Nuer belongs to the Dinka-Nuer group of languages within the Eastern Sudanic group of the Nilo-Saharan language family. Earlier, we added Dinka from the same subgroup, and these languages must be considered very close relatives. 4,782 lines of the 6,295 lines in the training data were identified as Dinka. Quite a large portion of those had a confidence score higher than 1.0. There were also 12 English-identified lines with clear English incursions. One of the development lines was also identified as English, which it mostly was. We filtered English out of both sets and lines with confidence scores equal to or higher than 1.0

[19] https://www.ethnologue.com/subgroup/39/

from the training set. Without confidence thresholds, Nuer attracted one false positive identification. Using the confidence thresholds took care of the single error, and Nuer received a perfect F-score.

7. Evaluation

So far, we have used only the development part of the FLORES-200 dataset to generate more internal test data for the HeLI-OTS language identifier. In this Section, we evaluate HeLI-OTS using the test partitions of the FLORES-200 dataset. During this research, we have not taken a look at the test set, and even though it is of high quality, it could very well include lines that we would consider to be multilingual. After adding the 20 languages, 113 of the 220 languages within the HeLI-OTS repertoire had corresponding ISO 639-3 identifiers in the FLORES-200 dataset.

7.1. Experiments with the Development Set

We started by identifying the development material for the 113 languages using HeLI-OTS with and without the confidence thresholds. With the thresholds, the Macro F1 on the development material reached 0.9907 and without 0.9973. The worst-performing language was Sango, which attained an F-score of only 0.2052 on the development set, while on the HeLI-OTS internal test set, it reached a perfect F-score. This signaled that there must be either a difference in the orthography used or a severe difference with the domain. From the 0.9990 F-score attained by Burchell et al. (2023) for Sango, it was clear that their training data was more similar to the FLORES-200 material than the one we have been using for HeLI-OTS. Without any understanding of the Sango language, it was not apparent what the mismatch was, but as HeLI-OTS allows additional models for languages, we decided to use the OpenLID training data to create a second model for Sango.

We treated the alternative Sango like the other languages in the previous two sections. The OpenLID data we use for training contained 255,491 lines of text for Sango, which we identified using the current HeLI-OTS models. Over 245,000 lines were identified either as Sango or left undetermined. The 450 lines identified as French in the training set were mostly consisting of only French words, so we filtered out all of them. The same was true for the 68 lines identified as English. We kept all the lines tagged either as undetermined or Sango, and from the rest, we took the lines with confidence scores of less than 1.0. For additional internal testing data for Sango, we took all the 997 lines of the FLORES-200 development set.

With confidence thresholds, the new Sango models attracted one false positive and reached the F-score of 0.9996 on the internal test set. Once the confidence thresholds were in use, Sango again attained a perfect F-score on the internal test set, which now also comprised the development data from the FLORES-200 dataset.

Next, we ran the HeLI-OTS again on the 113 language subset of the FLORES-200 development set. Now, Sango attained a perfect F-score with and without the confidence thresholds; the macro F1 score over all the languages rose to 0.9979 and 0.9984, respectively. The worst performing languages were now the Norwegian language pair with F-scores of 0.9633 for Bokmål and 0.9700 for Nynorsk. On the internal test set, they achieve 0.9814 and 0.9838, respectively. Using the OpenLID generated models, Burchell et al. (2023) attained 0.9719 and 0.9828. The largest mismatch was with 44 of the Nynorsk lines being identified as Bokmål. These lines seemed to be rather well-formed sentences in a Scandinavian language. We decided to take the opportunity to improve the HeLI-OTS Norwegian discrimination capability and created an alternate model for Nynorsk.

OpenLID training data for Nynorsk contained 101,140 lines of text, of which 73,501 were identified as Nynorsk and 15,486 as Bokmål. Swedish was next with 2,671 lines, and 2,525 lines were left undetermined. The lines left undetermined seemed to be of very poor quality, and the 1,282 lines identified as English contained English words. We left those two out and took all the Nynorsk lines and those with less than 1.0 confidence from the ones identified to be written in other languages. We were left with 96,116 lines of new training data for Nynorsk. We also added all the development lines from FLORES-200 to the internal test set.

Adding the alternative Nynorsk model made the results slightly worse on both the FLORES-200 development set and our internal test set, so we decided to roll HeLI-OTS back to having only one model for Nynorsk. FastText is a discriminative classifier, and this might be why its performance is better on this close language pair than a generative classifier like HeLI-OTS.

We also decided that these experiments on the FLORES-200 development set would now be finished and set out to evaluate the system on the test partition.

7.2. Final Results

On the development set, the results with confidence thresholds were better as the macro F1 over all the 220 languages was 0.9401 vs. 0.9042. However, the macro F1 scores over the 113 relevant languages were better without the confidence thresholds: 0.9984 vs. 0.9979. The same situation pre-

vailed on the test set with almost identical figures for the relevant language F-score. Without confidence thresholds, HeLI-OTS attained a 0.9985 F-score on the relevant languages and 0.9223 over all the 220 languages in its repertoire. With the thresholds, the F-score on the 113 languages was 0.9979, and for all 220, it was 0.9446.

Even though it is not directly indicated, it seems that the results described by Burchell et al. (2023) are macro averages over the relevant languages. When calculated from the language level results presented in the article, the macro average F1 over the 113 relevant languages is 0.9904 for OpenLID models and 0.9815 for the NLLB models. The selected language repertoire favors HeLI-OTS and especially the OpenLID over the NLLB models, as we added languages based on how well OpenLID had fared on this very test set. However, Burchell et al. (2023) showed that OpenLID was overall more accurate than the NLLB. With the 113 languages we have examined here, the results of the HeLI-OTS are more than four times closer to a perfect F1 score than the OpenLID models and more than eight times closer than the NLLB.

8. Discussion and Conclusions

Our goal was to integrate new languages into the HeLI-OTS language repertoire with minimal negative effects on the accuracy of the existing 200 languages.

Without the use of confidence thresholds, the 20 added languages attract 11 false negatives and 63 false positives, which average 0.6 and 3.2, respectively, per language. For all the 220 languages, the corresponding figures are 19.5 for both per language.

The macro F1 score on the internal dataset was 0.9961 over the 200 languages, and after adding 20 new languages, some with close relatives in the original repertoire, the macro F-score over the 220 languages was 0.9963.

These two measures, together with the excellent evaluation results using the FLORES-200 test set, show that we were able to accommodate new languages without deteriorating the performance of the HeLI-OTS.

The HeLI-OTS version 2.0 includes language models described in this article and is now available for download from Zenodo (Tommi Jauhiainen and Valosaari, 2024).[20]

[20] http://urn.fi/urn:nbn:fi:lb-2024040301

9. Acknowledgements

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under grant number 358720 (FIN-CLARIAH – Developing a Common RI for CLARIAH Finland).

10. Bibliographical References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Amir Khusru Akhtar, Gadadhar Sahoo, and Mohit Kumar. 2017. Digital corpus of santali language. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 934–938.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

William B. Cavnar and John M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

Ute Dieckmann, Mietta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen, and Krister Linden. 2023. The pipeline for publishing resources in the language bank of finland. In *Selected Papers from the CLARIN Annual Conference 2022*, number 198 in Linköping Electronic Conference Proceedings, pages 33–43, Sweden. Linköping University Electronic Press. CLARIN Annual Conference ; Conference date: 10-10-2022 Through 12-10-2022.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. Ethnologue: Languages of the world. http://www.ethnologue.com. Online version.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

Hossein Hassani, Dzejla Medjedovic, et al. 2016. Automatic kurdish dialects identification. *Computer Science & Information Technology*, 6(2):61–78.

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

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022a. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Linden. 2023. Tuning heli-ots for guarani-spanish code switching analysis. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, CEUR Workshop Proceedings, Germany. CEUR-WS.org. Iberian Languages Evaluation Forum : IberLEF 2023, IberLEF 2023 ; Conference date: 26-09-2023 Through 26-09-2023.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2015. Language Set Identification in Noisy Synthetic Multilingual Documents. In *Proceedings of the Computational Linguistics and Intelligent Text Processing 16th International Conference (CICLing 2015)*, pages 633–643, Cairo, Egypt.

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. The COLING 2016 Organizing Committee.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017. Evaluation of language identification methods using 285 languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 183–191, Gothenburg, Sweden. Association for Computational Linguistics.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019a. Language model adaptation for language and dialect identification of text. *Natural Language Engineering*, 25(5):561–583.

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.

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, and Krister Lindén. 2022b. Language identification as part of the text corpus creation pipeline at the Language Bank of Finland. In *The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, pages 251–259, Uppsala, Sweden.

Tommi Jauhiainen, Marcos Zampieri, Timothy C Baldwin, and Krister Lindén. 2024. *Automatic Language Identification in Texts*. Synthesis Lectures on Human Language Technologies. Springer, United States.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In

Proceedings of the ACL 2012 System Demonstrations, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Gertjan van Noord. 1997. *TextCat*. Software available at http://odur.let.rug.nl/v̈annoord/TextCat/.

11. Language Resource References

Tommi Jauhiainen and Heidi Jauhiainen. 2023. *HeLI-OTS 1.5*. University of Helsinki.

Tommi Jauhiainen, Heidi Jauhiainen, and Santtu Valosaari. 2024. *HeLI-OTS 2.0*. University of Helsinki.