Language Identification as part of the Text Corpus Creation Pipeline at the Language Bank of Finland

Tommi Jauhiainen[a], Jussi Piitulainen[a], Erik Axelson[a] and Krister Lindén[a]

[a]*University of Helsinki, Finland*

Abstract

The Language Bank of Finland hosts text corpora originating from Finland. Two of the most used ones are the Newspaper and Periodical Corpus of the National Library of Finland and the Suomi24 Corpus. The Language Bank has received considerable additions to both corpora and is currently creating new versions of the corpora. We are debuting language identification as part of the corpus creation pipeline. As a language identifier, we are using our recently published HeLI-OTS software. This paper investigates the results and the quality of language identification.

We created a new dataset for evaluating the efficacy of language identification by extracting random samples from both corpora and manually annotating their language. We were especially interested in seeing how the relatively low OCR quality of the oldest part of the KLK-fi collection will affect language identification when using an off-the-shelf program like HeLI-OTS. The oldest part of the KLK-fi collection and many parts of the Suomi24 corpus contain Finnish written dialectally or otherwise differing from the standard written Finnish. HeLI-OTS software includes several separate language models for dialectal Finnish, and in this article, we evaluate their usefulness using the new data set. For both corpora, the overall micro F1 scores increase when using the additional dialectal models. Additionally, we take a detailed look at some language identification errors and discuss possible solutions.

Keywords

text corpus, language identification, OCR

1. Introduction

The Language Bank of Finland[1] (LBF) is a virtual research infrastructure maintained by the FIN-CLARIN cooperation. The LBF hosts text corpora in Finnish as well as in other languages if they originate from or are used in Finland. Two of LBFs most used corpora are the Finnish Subcorpus of the Newspaper and Periodical Corpus of the National Library of Finland (KLK-fi)[2] and

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15–18, 2022

tommi.jauhiainen@helsinki.fi (T. Jauhiainen); jussi.piitulainen@helsinki.fi (J. Piitulainen); erik.axelson@helsinki.fi (E. Axelson); krister.linden@helsinki.fi (K. Lindén)
https://researchportal.helsinki.fi/en/persons/tommi-jauhiainen (T. Jauhiainen); https://researchportal.helsinki.fi/en/persons/jussi-piitulainen (J. Piitulainen); https://researchportal.helsinki.fi/en/persons/erik-axelson (E. Axelson); https://researchportal.helsinki.fi/en/persons/krister-linden (K. Lindén)
0000-0002-6474-3570 (T. Jauhiainen); 0000-0002-3835-2189 (J. Piitulainen); 0000-0002-7755-412X (E. Axelson); 0000-0003-2337-303X (K. Lindén)

[1] https://www.kielipankki.fi/language-bank/
[2] http://urn.fi/urn:nbn:fi:lb-2016050302

the Suomi24 Sentences Corpus 2001-2017 (suomi24-2001-2017)[3]. The over 5 billion tokens of the KLK-fi corpus originate from the Finnish magazines and newspapers starting from 1820, which the National Library of Finland (NLF) has digitized. The suomi24-2001-2017 corpus contains over 4 billion tokens from the various discussion forums of the Suomi24 social networking website from 2001 to 2017. LBF has received considerable additions to both corpora: c. 0.5 billion tokens for Suomi24 and c. 18 billion tokens for NLF. The first version of the new Suomi24 corpus is already available[4], and new versions of the KLK corpora are planned to be published during summer 2022.[5] We are debuting language identification (LI) as part of this process. The LBF has previously participated in using LI in creating text corpora for underresourced Uralic languages within the Finno-Ugric Languages and the Internet -project [1].

As a language identifier, we are using the recently published HeLI-OTS software.[6] HeLI-OTS is an off-the-shelf language identifier equipped with language models for 200 languages. In HeLI-OTS, the ISO 639-3 standard is used to determine the language division, and when determining the language of a text, the program returns an ISO 639-3 identifier as a result. As a method for LI, the program uses the HeLI method developed at the University of Helsinki [2]. This method is currently state-of-the-art for LI between many languages, as evidenced by its first place on the ULI-178 track of the Uralic Language Identification (ULI) 2020 and 2021 shared tasks [3]. We are identifying the language of each sentence, and it will be possible to use the resulting information as part of queries in the Korp user interface.[7] This paper investigates the results and the quality of the LI process.

In Section 2, we explain how sentences are extracted from the files received from the NLF. Section 3 describes the evaluation dataset created from the new NLF and Suomi24 material. The following Section provides the numerical results of the LI on this dataset. In Section 5, we take a closer look at the errors the best models made on the development sets of both materials. The last Section discusses some improvement ideas and concludes the paper's findings.

2. Tokenization and sentence detection

Many of the "sentences" problematic for LI in our corpora are not well-formed sentences in the first place. Therefore, we thought it essential to shed some light on the process of how these sentence-like objects were constructed. The rest of the paper will refer to these objects as sentences even if they could not be considered sentences by any linguist.

Each page of the NLF magazines and newspapers of the corpus has been OCR'd into an XML file in ALTO/METS format[8]. The XML file consists of one <*Page*> element consisting of <*TextBlock*> elements that consist of <*TextLine*> elements consisting of <*String*> elements. One Page comprises one page. One TextBlock comprises one block of text, i.e., paragraph, title, table, legend, and so on. One *TextLine* usually contains one line separated by a line/page break. *String* is basically anything separated by space. Each *TextBlock* is retokenized with the command-line

[3] suomi24-2001-2017-korp-v1-1, http://urn.fi/urn:nbn:fi:lb-2020021803

[4] http://urn.fi/urn:nbn:fi:lb-2021101525

[5] http://urn.fi/urn:nbn:fi:lb-202009152

[6] http://urn.fi/urn:nbn:fi:lb-2022021303

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

[8] https://www.loc.gov/standards/alto; https://www.loc.gov/METS/

tool *finnish-tokenize* (part of *Finnish Tagtools* v1.5.1[9]). The input to *finnish-tokenize* is one string that contains all *TextLine* and *String* elements in the given *TextBlock*, separated by spaces. The result from *finnish-tokenize* is a list of sentences, where a sentence is a list of tokens. Tokenizations from the XML file and *finnish-tokenize* are aligned, preferring the result from *finnish-tokenize*. Alignment is needed as the XML file contains attributes such as OCR character confidence and word confidence values for each token. We used *finnish-tokenize* for the texts written in other languages as well, which is not optimal, but remedying this is out of the scope of this article.

For the Suomi24 material, we used UDPipe version 1.2.0 [4] with the *finnish-tdt* (Turku Dependency Treebank) model [5] to segment paragraphs into sentences and tokens.

3. Evaluation environment

In order to improve and test the efficiency of the language identifier, we manually annotated random samples from the corpora. As both of the corpora contain fairly recent material, we scrambled all the numbers within the sentences to anonymize any possible personal information they might contain, such as street addresses and phone numbers. The scrambled texts contain the same amount of number characters as the originals, and each of the original number characters has been replaced by a random number character. For example, “Salpausseläntie 7 E 23 , 00710 Helsinki 71 , puh. 225 932 .” could become “Salpausseläntie 2 E 57 , 55667 Helsinki 06 , puh. 116 071 .”. The development and the test data are available via LBF[10] with a CC-BY-NC license for the Suomi24 part and CC-BY license for the NLF part.

3.1. NLF Random evaluation set

The LBF has published four different types of corpora originating from the material collected by the NLF.[11] Two of these are language-based, and two are based on the date of the publication of the material in question. So far, the material for the language-based, Swedish or Finnish, has been selected using the language tags provided by the NLF. In 2021, LBF received a new batch of material from the NLF.

The random development and test sets for the NLF data were generated from publications OCRd by the NLF during 1.1.2014-30.6.2014 and 1.1.2015-30.6.2015.[12] The publications contained “texts” tagged as written in Finnish, Swedish, English, and German. The number of texts in each language can be seen in Table 1. We randomly selected 500 for both Finnish and Swedish, 200 for English, and all six German texts from those texts. After this, from each text, we randomly sampled 20 “sentences” generated as described in Section 2. The number of sentences for each language is also seen in Table 1.

We manually identified the language of each sentence, except those of the Russian collection, due to our low ability to read Cyrillic. Cyrillic sentences in the collections of other languages were annotated as Russian if they contained more than individual Cyrillic letters. This means

[9] http://urn.fi/urn:nbn:fi:lb-2021042101

[10] http://urn.fi/urn:nbn:fi:lb-2022021301

[11] http://urn.fi/urn:nbn:fi:lb-2021092404

[12] They contain publications from the early 1800s to the 2010s.

Table 1
Number of texts and sampled sentences by language indicated in the NLF collection.

Language	# texts	# sentences	# sampled sentences
Finnish	18,897	697,616	10,000
Swedish	6,232	704,143	10,000
English	475	88,000	4,000
German	6	40,746	120

Table 2
Actual languages observed in the sentences from the NLF based on manual inspection and the number of sentences in each language in the development and the test sets.

Language	Finnish	Swedish	English	German	Development set	Test set
Finnish (fin)	8,204	53	144	62	4,394	4,069
Swedish (swe)	73	7,804		11	4,060	3,828
English (eng)	92	11	3,489	4	1,856	1,740
German (deu)	30	41		2	43	30
Unknown (xxx)	1,272	1,783	363	33	1,683	1,768
Russian (rus)	319	285			0	604
French (fra)	1	4		6	2	9
Latin (lat)		4			2	2
Italian (ita)		3			2	1
Dutch (nld)		1			1	0
Multilingual	9	11	4	2	17	9

that sentences annotated as "rus" could be written in any language using Cyrillic characters. If the sentence seemed to consist solely of one or more proper names, it was annotated as "xxx". The annotation "xxx" was also used if it was impossible to determine the language because the sentence contained only non-linguistic contents such as severe OCR errors or numbers. The number of sentences in each detected language can be seen on the left side of Table 2. 26 sentences were also deemed multilingual.

From these sentence collections, we selected the first 50% of the sentences from each of the four NLF tagged languages for the development set and the last 50% for the test set. The language distribution within the two sets was not equal, as can be seen on the right side of Table 2. For example, all the Cyrillic (rus) texts were in the latter half of the original collection and thus ended up in the test set. When LI methods are developed using these sentence collections, it is essential that the information about the language distribution in the test set is not used to improve the results.

3.2. Suomi24 Random evaluation set

The new material for the Suomi24 corpus was already imported to Korp, so we used the Korp interface concordance function to sample 3,000 sentences from 2018–2020 randomly. The total number of sentences for each year ranged from 12.5 to 14.5 million. Unlike in the NLF material, there are no existing language tags available for the texts. The supposition was that most of

Table 3
Languages observed in sampled sentences from Suomi24 and the number of sentences in each language in the development and the test sets.

Language	All	Development set	Test set
Finnish (fin)	8,672	4,326	4,346
Swedish (swe)	26	11	15
English (eng)	22	9	13
Unknown (xxx)	271	149	122
Arabic (ara)	2	0	2
Multilingual	7	5	2

the texts would be in Finnish, but texts or sentences in other languages could also be found. The language of these 9,000 sentences was annotated manually, and it supported the prior supposition as only c. 0.6% of the sentences were written in languages other than Finnish. Again, we divided the dataset into equal portions for development and testing. The dataset contained some Swedish and English sentences, but over 96% of the dataset was clearly in Finnish, as can be seen in Table 3.

4. Results

In this Section, we use the HeLI-OTS language identifier to identify the language of the sentences in the evaluation set. The 1.1 version of the identifier includes only one language model for Finnish trained from one million sentences of 2002 Web crawl data from Leipzig Corpora Collection [6]. The 1.2 and 1.3 versions have several additional language models for Finnish, which were created from data gathered during the Finno-Ugric Languages and the Internet -project to overcome the problem of dialectal Finnish being identified as some of the close relatives of Finnish such as Kven Finnish or Karelian. This same problem had been noticed by early users of the 1.1 version, which is why the 1.2 version was developed. The efficacy of adding the additional language models for Finnish was never adequately evaluated during the Finno-Ugric Languages and the Internet -project, and the current study describes the first proper evaluation of this strategy.

In these first experiments, we set a baseline for LI results for this dataset, focusing on the results attained on the development set. We evaluate the HeLI-OTS 1.3 version, with and without the additional Finnish language models. The results for the NLF development set are in Table 4 and the results for the Suomi24 development set are in Table 5. Using the additional models for Finnish, the overall micro F1 score for the NLF test set was 87.11 and 97.60 for the Suomi24 test set. The results on the test set could be improved easily by using the information in the development set in various ways, such as using the development set as additional training material. However, we wanted to see how the off-the-shelf version of the language identifier fares with both sets and leave further experiments to future work.

Table 4
Results of the HeLI-OTS 1.3 with (w) and without (wo) additional language models for Finnish on the NLF development set.

Language	wo recall	wo precision	wo F1	w recall	w precision	w F1
Finnish (fin)	88.71	97.43	92.86	91.35	95.41	**93.34**
Swedish (swe)	77.86	96.84	**86.32**	77.78	96.93	86.31
English (eng)	91.86	97.71	**94.70**	91.81	97.65	94.64
German (deu)	72.09	45.59	**55.86**	72.09	44.93	55.36
Unknown (xxx)	28.70	99.18	**44.52**	28.70	99.18	**44.52**
French (fra)	100.0	8.33	**15.38**	100.0	8.0	14.81
Latin (lat)	100.0	5.56	**10.53**	100.0	5.56	**10.53**
Italian (ita)	50.0	8.33	**14.29**	50.0	7.14	12.50
Dutch (nld)	100.0	3.45	**6.67**	100.0	3.45	**6.67**
Multilingual	0.0	100.0	**0.0**	0.0	100.0	**0.0**
All	76.98	96.05	85.46	77.91	95.20	**85.69**

Table 5
Results of the HeLI-OTS 1.3 with (w) and without (wo) additional language models for Finnish on the Suomi24 development set.

Language	wo recall	wo precision	wo F1	w recall	w precision	w F1
Finnish (fin)	95.95	99.66	97.77	98.27	99.23	**98.75**
Swedish (swe)	90.91	83.33	**86.96**	90.91	83.33	**86.96**
English (eng)	44.44	20.0	27.59	44.44	21.05	**28.57**
Unknown (xxx)	34.23	100.0	**51.0**	34.23	100.0	**51.0**
Multilingual	0.0	100.0	**0.0**	0.0	100.0	**0.0**
All	93.69	99.25	96.39	95.91	98.85	**97.36**

5. Error analysis

In this Section, we focus on the errors made by the HeLI-OTS language identifier with additional dialectal language models for Finnish. In this study, we refrain from studying the errors made on the test set in order to be able to continue developing the LI methods using the collection. Therefore we focus on analyzing some of the errors made on the development set in detail.

5.1. KLK

From the four languages of interest, the lowest F1 score was attained for German with both relatively low recall and low precision. From the 43 sentences annotated as German in the development set, only 31 were identified as such. The recall errors seem to be primarily due to the combined effect of OCR errors and the HeLI language repertoire, including several closely related languages: Swiss German (4 errors), Bavarian (3), and Pfaelzisch (2). An example of a sentence identified as Swiss German is "Die Ver n i chtung der boi sch ewistischen Ge f ahr und der plutokratischen Ausbeutung wird die Möglichkeit einer fr i edlichen harmonischen und f r uchtbaren Z us amm enar b eit aller Volker des europäischen Kontinents so w ohi auf p oli tisch em als auf wirtschaftlichem und kultur eli em Geb iet scha f f e n. »]". Most of the precision errors for German came from short sentences tagged as Swedish. An example of such behaviour is: "Hummern hade galt öfwer bord .", where the misclassification is probably due to the combination of OCR errors, "galt" should be "gått", and not using current spelling, "öfver" is "över" in standard spelling.

The other three languages all had F1 scores over 85, with Swedish having the lowest score due to a very low recall of 77.78%. Of the 902 erroneously identified Swedish sentences, 87 were identified as Danish, 47 as Norwegian Bokmål, 38 as Nynorsk, 36 as Lushai, 35 as Swiss German, 33 as Finnish, and 30 as Kölsch. The rest were divided between 103 different languages. The sentences incorrectly identified as Danish seem to be primarily due to OCR errors such as "6 anlebning af OfebaftionenS od) målaren ©parfS i fenafte nummer af SÖ .", which should be a Swedish sentence: "I anledning af Redaktionens och målaren Sparss i senaste nummer af B.". OCR errors are to blame also for both of the Norwegian languages, Swiss-German, and Kölsch. The samples mistaken as Finnish are primarily due to Finnish proper names. The samples mistaken as Lushai are due to "tel" being a word in Lushai and the abbreviation for telephone number "tel." being the only word in 30 Swedish sentences.

The version with extra language models for Finnish made significantly more precision errors for Finnish, the number of errors rising from 103 to 193. The gold-standard language for 142 of these erroneous identifications was tagged as "xxx", meaning that they were either non-lingual or consisted of names or numbers. The 142 sentences included Finnish personal names "Heikki Jussila .", addresses "Iso Uusikatu N:o 29 .", place names "HELSINKI-TAMPERE", and names of companies "KONE O.Y .", as well as their combinations. They also included some incomprehensible sentences, probably due to OCR problems, which might have originated from Finnish such as "§nraästi !" and "Miwälista .". Examples of non-lingual sentences in the development set are "m . , y .", "g .", and "1 § .".

Of the 152 incorrectly identified English lines, only two were actual sentences containing more than one word. Both "Cornelius was obedient ." and "In Jesus ' name ." were identified as Afrikaans.

5.2. Suomi24

On the Suomi24 development set, the worst-performing language was English, with an F score of only 28.57. Only four of the nine sentences tagged as English were well-formed sentences, all identified as English. For example, two sentences, "wtf" and "WTF?" were identified as Lushai. These kinds of errors lowered the recall for English. The precision for English was down due to 15 sentences tagged as something else identified as English. Five of the fifteen were HTTPS addresses containing English words, two were Finnish English multilingual, one was incorrectly labelled as Finnish, and most of the others contained proper names.

The Finnish sentences were most often identified as one of the close languages within the repertoire. Of the 75 incorrectly identified Finnish sentences, eight were identified as Karelian, six as Ingrian, and five as Estonian. Some of the sentences identified as Karelian were reasonably well-formed Finnish sentences, such as "En kieliä vieroksu." and "Ei eletä.". Also, some of the sentences identified as Ingrian were well-formed Finnish sentences such as "Kolmestihan se sinne kaislikkoon lensi.". As the Ingrian training corpus contained only 582 words, any word seen in the Ingrian training corpus is also heavily scored as Ingrian in the mystery text. For example, the word “sinne” is used in well-formed Ingrian sentences in the training corpus.

6. Discussion and Conclusions

In many cases where the language of a sentence was incorrectly predicted, the correct language had almost as good a score as the predicted language. It could be beneficial to give some of the commonly expected languages a slight increase to the prior probability. These probabilities could be determined using the development set for this particular evaluation setting. However, they would not be suitable as ready-to-use general-purpose probabilities for the HeLI-OTS software.

Another way to improve the results on the test set would be to restrict the HeLI-OTS language repertoire based on errors seen in the development set. For example, none of the sentences identified as written in one of the languages related to Finnish in the Suomi24 development set seemed to be written in them. Leaving them out of the repertoire when processing the test set would undoubtedly increase the accuracy of the identifications.

Many sentence identification errors were due to the presence of named entities using common words from other languages. An example of such a sentence is “Allmänna Svenska Elektriska Aktiebolaget, Vesteräs, Ruotsi : laite aikalisiin akkumulaattoreihin.”, which contains a name of a Swedish company in a Finnish sentence.

In addition to using the development material as in-domain training material when processing the test set, there are ways to adapt the language models to the material being identified. Several international LI competitions have been won using adaptive language models with the HeLI and Naive Bayes methods [7, 8]. This functionality will be incorporated into HeLI-OTS in the future, and we will also evaluate its efficacy on this dataset.

Especially the NLF corpus contains many non-lingual sentences, which could be identified as such if an unseen language detection capability were incorporated into the software. In many cases, the OCR process had created extra spaces inside words. HeLI-OTS uses spaces in word tokenization and gives each word equal value, which leads to parts of words being treated as words and thus into erroneous identifications. It would be worthwhile to see whether it would be beneficial to ignore spacing completely in the NLF material or try another word tokenization strategy.

Our results indicate that implementing some of the previously mentioned functionalities will significantly improve the accuracy of language identification.

Acknowledgments

The research presented in this article has been funded by the Finnish Research Impact Foundation[13] from its Tandem Industry Academy 2020 call in cooperation with Lingsoft Oy[14], the Academy of Finland, and the University of Helsinki.

[13] https://www.vaikuttavuussaatio.fi/en/

[14] https://www.lingsoft.fi/en

References

[1] H. Jauhiainen, T. Jauhiainen, K. Lindén, Building Web Corpora for Minority Languages, in: Proceedings of the 12th Web as Corpus Workshop, European Language Resources Association, Marseille, France, 2020, pp. 23–32. URL: https://www.aclweb.org/anthology/2020.wac-1.4.

[2] T. Jauhiainen, K. Lindén, H. Jauhiainen, HeLI, a word-based backoff method for language identification, in: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 153–162. URL: https://www.aclweb.org/anthology/W16-4820.

[3] B. R. Chakravarthi, M. Gaman, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, M. Zampieri, Findings of the VarDial evaluation campaign 2021, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Kyiv, Ukraine, 2021, pp. 1–11. URL: https://www.aclweb.org/anthology/2021.vardial-1.1.

[4] M. Straka, J. Straková, Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, in: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 88–99. URL: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

[5] K. Haverinen, J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski, F. Ginter, Building the essential resources for Finnish: the Turku Dependency Treebank, Language Resources and Evaluation 48 (2014) 493–531.

[6] D. Goldhahn, T. Eckart, U. Quasthoff, Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 759–765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

[7] T. Jauhiainen, H. Jauhiainen, K. Lindén, HeLI-based experiments in Swiss German dialect identification, in: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 254–262. URL: https://www.aclweb.org/anthology/W18-3929.

[8] T. Jauhiainen, H. Jauhiainen, K. Lindén, Naive Bayes-based experiments in Romanian dialect identification, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Kyiv, Ukraine, 2021, pp. 76–83. URL: https://www.aclweb.org/anthology/2021.vardial-1.9.