Tuning HeLI-OTS for Guarani-Spanish Code Switching Analysis

Tommi Jauhiainen^{1,*}, Heidi Jauhiainen¹ and Krister Lindén¹

¹Department of Digital Humanities, University of Helsinki, Helsinki, Finland

Abstract

This article describes a system created for the first subtask of the GUA-SPA - Guarani-Spanish Code Switching Analysis shared task held as part of the IberLEF 2023 evaluation campaign. The system was based on the HeLI-OTS off-the-shelf language identifier and ad-hoc rules for detecting named entities, unknown languages, other tokens, and words with mixed Spanish Guarani language. With our system, we attained the second position in the subtask with an F-score of 0.9139.

Keywords

Language identification, Code-switching, Named entity recognition

1. Introduction

The Guarani-Spanish Code Switching Analysis (GUA-SPA) shared task [1] was held as part of the IberLEF 2023 evaluation campaign [2]. The IberLEF evaluation campaigns have contained a large number of shared tasks related to various Iberian languages, but the GUA-SPA is the first language identification-related shared task so far[3, 4]. In addition to identifying individual words as either Guarani or Spanish, the task included separate classes for words mixing both languages, words in other languages, other tokens, and named entities. We entered the shared task determined to make use of our previously published off-the-shelf language identifier HeLI-OTS [5]. HeLI-OTS has not been evaluated on word-level language identification tasks before, nor does it currently include the possibility for language set identification [6] or for text segmentation by language [7]. HeLI-OTS also always labels words in named entities with the language they are most likely found in the training data for the language identifier. The same goes for non-alphabetic characters like commas and dots if they appear connected to alphabetic characters. However, given only non-alphabetic characters, HeLI-OTS already gives the output language as "und", which is the official ISO 639-3 identifier for any undetermined language. We set out to build the missing functionalities on top of the existing software

IberLEF 2023, September 2023, Jaén, Spain

^{*}Corresponding author.

^{☆ {}firstname.lastname}@helsinki.fi (T. Jauhiainen); {firstname.lastname}@helsinki.fi (H. Jauhiainen); {firstname.lastname}@helsinki.fi (K. Lindén)

https://researchportal.helsinki.fi/en/persons/tommi-jauhiainen (T. Jauhiainen);

https://researchportal.helsinki.fi/en/persons/heidi-jauhiainen (H. Jauhiainen);

https://researchportal.helsinki.fi/en/persons/krister-linden (K. Lindén)

CEUR Workshop Proceedings (CEUR-WS.org)

using the GUA-SPA training data as our guide. HeLI-OTS is available from Zenodo: https://zenodo.org/record/7066611.

In Section 2, we introduce some previous work for language identification for the Guarani language as well as for word-level language identification for code-switching. Section 3 contains the descriptions of the evaluation setting of the shared task as well as the corpora used. Section 4 introduces HeLI-OTS, an off-the-shelf language identifier for text which is used as the underlying language identification system. In Section 5, we describe some of the more interesting experiments we conducted when participating in the shared task. Section 6 is a description of the whole system pipeline for our best submission focusing on the ad-hoc rules on top of the HeLI-OTS language identification functionality. In Section 7, we go over the final results, and in the last Section, we discuss some of the challenges and ideas for future work regarding the task at hand.

2. Previous Work

Most language identification systems aim to identify the language of sentences or longer passages of texts. For a detailed overview of the research in language identification and some related tasks, we refer the reader to a comprehensive survey by Jauhiainen et al. [8]. In code-switching analysis, language identification on the word level is needed.

2.1. Language Identification for the Guarani Language

In the ISO 639-3 standard, Guarani (grn, https://iso639-3.sil.org/code/grn) is considered a macro language containing five individual languages: Western Bolivian Guaraní (gnw), Paraguayan Guaraní (gug), Eastern Bolivian Guaraní (gui), Mbyá Guaraní (gun), and Chiripá (nhd). From the introduction to the GUA-SPA shared task, it seems that the language concerned in the task might be the individual Paraguayan Guaraní (https://codalab.lisn.upsaclay.fr/competitions/ 11030#learn_the_details). As the HeLI-OTS language identifier already contained language models for the macro language and the shared task did not specify any ISO 639-3 codes, we did not do any research on the differences between individual Guarani languages.

Guarani language has been mentioned in relatively few language identification experiments in the past. For his Ph.D. thesis, Rodrigues [9] built a language identifier using Universal Declarations of Human Rights (UDHR) as the training material. Paraguayan Guaraní was one among the 371 other languages in the repertoire of the system. Later, using the JRC-Acquis corpus [10] to evaluate the vector-space classifier introduced by Prager [11], Lui [12] notes that test instances written in Portuguese were erroneously identified as Guarani in his experiments. He believed that this was due to domain differences between Portuguese training and testing data. In the light of our recent experiences with the Guarani training corpora for HeLI-OTS, which we give some details in Section 5, we believe that his training data for Guarani could have been similarly saturated with Spanish vocabulary causing the observed misclassifications. When evaluating several language identification methods in hard contexts, we used Guarani Wikipedia as training material for our Paraguayan Guaraní models and the Guarani UDHR as the testing data among the 284 other languages [13]. Using the HeLI method [14], we attained over 99% recall and precision for Guarani for test texts of 40 characters or longer while the overall macro F1 score for all the 285 languages is 98,5% at 40 characters. Caswell et al. [15] train language identification models for 1,629 languages, and in their comparison of precision filtering approaches, they show their models for Guarani had 100% recall, but a precision of only 4.0%. Using precision filtering, they were able to boost the precision to 44.0% at the cost of recall dropping to 92.1%. Their notes for Guarani read "some lexical overlap with Spanish", which is undoubtedly due to a significant amount of code-switching in their training corpora.

Góngora et al. [16] describe the creation of a text corpus for the Guarani language. The corpus consists of two separate parts; the first is a crawled news corpus with parallel Guarani-Spanish sentences, and the second is a collection of monolingual tweets in Guarani. For the parallel corpus, they used a set of Guarani words to build a seed list of addresses containing text in Guarani. To identify the language of tweets, they first empirically inspected the results given by the Twitter API and noticed that there were no tweets identified as Guarani, even though there were tweets entirely written in Guarani. Using the corpus collected by Chiruzzo et al. [17] as training data, they trained a character 5-gram Naïve Bayes identifier for distinguishing between Guarani and Spanish. The identifier gained a very high accuracy on their test partition, but it was still not good enough for correctly identifying the language of the tweets. Finally, they created two lists of frequent words unique to Guarani and used them to identify the language of the tweets by counting the matching words and using different thresholds for tweets originating from Paraguay and elsewhere. In a similar work to create a corpus of Guarani texts, Agüero-Torales et al. [18] used three off-the-shelf language identifiers to determine whether the texts were written in Guarani. The three tools were: polyglot (https://polyglot.readthedocs.io/en/latest/ Detection.html), fastText (https://fasttext.cc/docs/en/language-identification.html), and textcat (https://www.nltk.org/ modules/nltk/classify/textcat.html). From 2.1 million tweets, 5.300 were identified as Guarani by at least one of the three classifiers. After automatic identification, they manually inspected the 5,300 tweets, of which only 150 were actually Guarani-dominant.

As far as we are aware, apart from the experiments described by Góngora et al., there has not been any language identification development focusing on the Guarani language before the present GUA-SPA shared task.

2.2. Language Identification for Code-Switching

The GUA-SPA is not the first shared task focusing on code-switching. The First Shared Task on Language Identification in Code-Switched Data was organized in 2014 [19]. The second code-switching shared task was held in 2016 [20]. Even though the Guarani - Spanish pair has not been a focus of much language identification research so far, Spanish has featured with other languages, such as English, as was the case in the first code-switching shared task. The first ENG-SPA task was won by Bar and Dershowitz [21] using a LibSVM-based support vector machine (SVM) system with a second-degree polynomial kernel. They used a collection of features from a window containing two words before and two words after the word being identified. They attained the weighted F1 score of 0.940 over six categories, including named entities, etc., in a similar way to the task at hand. The ENG-SPA pair was featured again in the second shared task. This task was won by Shirvani et al. [22] using logistic regression to label tokens utilizing various combinations of 14 feature types, including POS-tags for English and Spanish, the output of a separate NER-tagger in addition to, for example, character n-grams

and dictionaries. They attained the token-level weighted F1 score of 0.973.

Some more recent work on code-switching include using SVM's on code-mixed Hindi-English and Urdu-English social media text [23], word-level language identification for code-switching detection for Austronesian languages [24] with the fastText off-the-shelf language identifier [25], using subword embeddings for code-switched Bangla-English social media texts [26], codeswitching identification for under-resourced languages [27], word-level language identification in social media using convolutional neural networks (CNNs) [28], code-switching detection for Kannada-English texts using transformers [29], and code-switching detection for 16th-century letters [30].

Hidayatullah et al. [31] have recently published a review on language identification of codemixed texts.

3. Evaluation Setting

In the first phase, the participants were provided with the gold-labeled training data and the unlabeled development data. In the testing phase, they were given the gold labels for the development data and the unlabeled test data.

On track one, the task was to classify each pre-tokenized word into one of six categories: Spanish (es), Guarani (gn), mixed Spanish and Guarani (mixed), other languages (foreign), named entities (ne), and other tokens (other). On track two, the task was to additionally label the named entities as belonging to one of three groups: person, location, or organization. On the third subtask, it was also necessary to indicate whether the Spanish words were part of longer Spanish texts or were they included in otherwise Guarani sentences.

The training data contained 1,140 lines of text tokenized into 19,003 tokens. The development and the test data both had 180 lines with 2,989 and 2,857 tokens, respectively. The distribution of these tokens between the six classes in the training and the development data can be seen in Table 1.

Token type	Training	Development
gn	7,698	1,241
es	5,058	812
other	3,220	456
ne	2,510	414
mix	388	52
foreign	129	14
total	19,003	2,989

Table 1

The size and contents of the training and the development datasets in the number of tokens.

The main scoring method was the weighted F1 measure. Also, accuracy, weighted precision and recall, as well as macro precision, recall, and F1 score, were calculated.

4. HeLI-OTS

HeLI-OTS is an off-the-shelf language identification tool first published in May 2021. The current version, 1.4, was published in September 2022 (https://zenodo.org/record/7066611). It is an implementation of the original HeLI algorithm first developed by Jauhiainen [32] for his master's thesis. The method was first called HeLI when it was properly described after gaining the shared first position in the Discriminating between Similar Languages (DSL) shared task in 2016 [14]. In a recent evaluation, Jauhiainen et al. [5] compare the HeLI-OTS and fastText off-the-shelf language identifiers and show that fastText favors the recall of common languages while HeLI-OTS reaches very high accuracy for all languages.

For longer texts, the HeLI-OTS uses a word-based scoring method where each word is given equal weight when determining the language of the whole text. For example, using this scoring, the word "the" will have equal weight in determining the language of the sentence as the word "international". Many language identification methods divide the texts into overlapping character n-grams, giving each character n-gram equal weight, which would, e.g., give the word "international" ten times more weight than the word "the" if they are divided into word-internal character trigrams. In a basic HeLI implementation, each word is identified independently of the surrounding words so that the character n-grams generated do not span over words, and this is also the case for HeLI-OTS.

HeLI-OTS has seven different language model types: whole words and character n-grams from one to six characters. When identifying the language of the word, the word-based models are checked first: if the word is found from any of the languages known by the identifier, the word models are used for all languages, and those languages not knowing the word receive a penalty score. If the word is unknown to all languages, the word is divided into the highest length n-grams known by the identifier, e.g., six grams. If any of the 6 grams generated from the word are known by any of the languages, they are used. If they are not found in any of the languages, the method backs off to using 5-grams and so on if needed until unigrams of characters. The scores used for words and character n-grams are negative logarithms of their relative frequencies in the training corpora for the identifier. When scoring the word using character n-grams, the score for the word is the average of the scored n-grams. Using the average has proven to give reasonably comparable scores between the models. For a more formal and detailed description of the HeLI method, we recommend taking a look at chapter three of the Ph.D. thesis of the first author [33].

5. Experiments

As we began our experiments with the training data of the shared task, it quickly became evident that our Guarani training data for the language identifier contained a huge amount of Guarani - Spanish code-switching. Using erroneous identifications of the shared task training data as our guide, we manually removed a lot of Spanish words and named entities from the HeLI-OTS training data for Guarani. The number of words in the training corpus went down almost 12,000 words or by 8%. Some of the words affected are listed in Table 2.

During the development phase, the participants were given the possibility to submit their

Word	HeLI-OTS 1.4	HeLI-OTS 1.5 (in development)
de	1,842	925
la	558	232
del	470	265
У	440	225
paraguay	326	223
san	280	200
nacional	216	17
el	188	92
juan	160	94
josé	152	89

Table 2

The frequency of some Spanish words before (HeLI-OTS 1.4) and after cleaning (1.5).

predictions on the development data and receive results using the official measures. Before submitting the results, we used the overall recall as the measure to improve.

When starting with the publicly available HeLI-OTS 1.4, the overall token-level recall was 55,57% on the training set. One of the largest single misclassifications seemed to be 340 named entities being mapped as Abkhazian (abk). The tokens misclassified were exclusively the "@USER" token. Why this word was mapped to Abkhazian, a language written in a Cyrillic character set, is due to the nature of the HeLI-OTS training corpus. Most of the training corpora were the ones used in the Uralic Language Identification (ULI) shared task [34]. The training data for many of the languages come from the Leipzig Corpora collection [35] (https: //wortschatz.uni-leipzig.de/en/download). The corpora for Uralic languages come from the Wanca 2016 corpora [36] and some additional languages from the corpora used to train the language identifier for the Finno-Ugric Languages and the Internet project's web crawling system [37]. The web page listing the sources for the latter tells us that the Abkhazian training corpus was basically a Wikipedia export (http://suki.ling.helsinki.fi/LILanguages.html). This is also the case for many other languages, and during our experiments for the GUA-SPA shared task, we spent a lot of time cleaning the training data for the HeLI-OTS language identifier. In the end, however, we ended up excluding the most troublesome languages from the language repertoire for the current shared task and continued improving the HeLI-OTS models separately.

We did a short experiment using a separate NER-tagger for Spanish that is available at the Hugging Face [38] (https://huggingface.co/flair/ner-spanish-large). However, it did not seem to be able to predict the named entities any better than the ad-hoc rules we had created so far, explained in detail in the next Section. There were differences in the NE predictions, but combining the output would not have been straightforward, and we left further experiments with third-party NER systems to future work.

6. Final System Description

As the basic building block of our system, we had the HeLI-OTS off-the-shelf language identifier. We started experimenting with version 1.4 and made some modifications, some of which are present in the final version. One of the main modifications was to restrict the list of languages to 64 out of the 200 languages known by the HeLI-OTS. Our experiments on the training data indicated that it was advantageous to map some of the 62 other languages to one of the other categories than "foreign". The final mappings can be seen in Table 3. Of the languages remaining to be mapped to "foreign", only English was a language using the Latin alphabet. The modifications for mapping were implemented directly into the HeLI-OTSs *identifyLanguage* function.

Language (ISO 639-3)	Mapped category	
Mirandese (mwl)	Spanish (es)	
Portuguese (por)	Spanish (es)	
Galego (glg)	Spanish (es)	
Ido (ido)	Spanish (es)	
Cheyenne (chy)	Guarani (gn)	
Panjabi (pan)	Named entity (ne)	
Komi-Permyak (koi)	Named entity (ne)	
Japanese (jpn)	Other (other)	
Gujarati (guj)	Other (other)	
Yakut (sah)	Other (other)	
Bulgarian (bul)	Other (other)	
Amharic (amh)	Other (other)	

Table 3

The mappings from certain language codes to categories other than "foreign".

Before passing the text to the language identifier, we created preprocessing functions which handled detecting "mixed" words, named entities, as well as some special cases.

First, the preprocessor checks the word against a list of special cases which are directly mapped to certain categories. These include words such as Spanish weekdays and months starting with capital letters mapped to Spanish as well as some consisting solely of Guarani word-endings such as "-kuéra", "-gui", or "kuérape" which were mapped to Guarani. The word "@USER" was mapped to the named entity and some words such as "URL", "xd", and "Com" to the other category. If, after removing all digits, the word consisted only of "G", "a", or "-pe", it was mapped to the other category as well.

For detecting the named entities, we implemented a counter that indicated the position of the word in the sentence. It was reset after each line as well as if the previous word was either "@USER" or "URL", or in case it consisted only of punctuation characters. All the words beginning with an uppercase letter and followed by a lowercase letter that were not in the first position were marked as named entities. Also, if a word not in the first position was preceded by a named entity and followed by a word beginning with a capital letter, it was marked as a named entity if it was one of the words "de", "del", or "y". For detecting named entities in the first position, we generated lists of words which were mostly starting with lowercase letters in our training corpora for Spanish and Guarani. If a word not in these lists was found in the first position and beginning with a capital letter, it was marked as a named entity.

Mixed Spanish Guarani words were detected firstly by looking for Guarani endings indicating mixing: "-pe", "-kuéra", "-gua", "-kuérape", and "-guava". Secondly, if the word started with "oñe" and the rest of the word was identified as Spanish, it was marked as "mixed". Also, if

a word started with "o" and was not found in any of the language models of any of the languages, and the rest of the word was identified as Spanish, it was marked as mixed.

For the final system, we also concatenated the words annotated as Guarani or Spanish in the training and the development data to the training data for the respective languages.

7. Results

Our final run with the test data resulted in the scores seen in Table 4. The last modification of adding the words from the GUA-SPA training and development data to the training data for the language identifier improved the weighted F1 score from 0.9098 to 0.9139, which was our best result, 0.0242 behind the best results reached by the winning team.

Measure	score
Accuracy	0.9146
Weighted Precision	0.9160
Weighted Recall	0.9146
Weighted F1	0.9139
Macro Precision	0.7519
Macro Recall	0.7422
Macro F1	0.7244

Table 4

Our best results on the test set for task one.

8. Discussion and Conclusions

In the final phase of our experiments with the development data, the label pair with the most errors was named entities and Spanish. Forty-seven words tagged as named entities were identified as Spanish, and 45 words annotated as Spanish were identified as named entities. It would be worthwhile to continue the experiments by combining a specially trained Spanish NER tagger into the pipeline.

Some Spanish words were still identified as Guarani, and undoubtedly further cleaning of the Guarani training corpus could have improved the situation.

Combining HeLI-OTS language identification results with other word-level features using a classifier such as CRF could lead to better results than using the ad-hoc rules we generated for the shared task, but experimenting with them was beyond the scope of our participation for this shared task.

Our results are only slightly behind the results of the winning team, but we consider it a good run in light of our ad-hoc rules being relatively simple when compared with, for example, the ones used by the winners of the previous code-switching shared tasks. We also consider our participation a success as it pointed us toward some problems in the training data for our HeLI-OTS language identifier. We were able to improve the quality of the HeLI-OTS models for the two languages as well as for many others, which included considerable amounts of Spanish text. These improvements will be part of the forthcoming version of the HeLI-OTS.

Acknowledgments

We thank the task organizers for their work. Writing this article has been supported by the Academy of Finland (Funding decision no. 341798).

References

- L. Chiruzzo, M. Agüero-Torales, G. Giménez-Lugo, A. Alvarez, Y. Rodríguez, S. Góngora, T. Solorio, Overview of GUA-SPA at IberLEF 2023: Guarani-Spanish Code-Switching Analysis, Procesamiento del Lenguaje Natural 71 (2023).
- [2] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.
- [3] J. Gonzalo, M. Montes-y Gómez, P. Rosso, Iberlef 2021 overview: Natural language processing for iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop, 2021, pp. 1–15.
- [4] J. Gonzalo, M. Montes-y Gómez, F. Rangel, Overview of iberlef 2022: Natural language processing challenges for spanish and other iberian languages (2022).
- [5] T. Jauhiainen, H. Jauhiainen, K. Lindén, HeLI-OTS, off-the-shelf language identifier for text, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3912–3922. URL: https://aclanthology.org/2022.lrec-1.416.
- [6] T. Jauhiainen, K. Lindén, H. Jauhiainen, Language Set Identification in Noisy Synthetic Multilingual Documents, in: Proceedings of the Computational Linguistics and Intelligent Text Processing 16th International Conference (CICLing 2015), Cairo, Egypt, 2015, pp. 633–643.
- [7] H. Yamaguchi, K. Tanaka-Ishii, Text Segmentation by Language Using Minimum Description Length, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 969–978. URL: https://aclanthology.org/P12-1102.
- [8] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, K. Lindén, Automatic Language Identification in Texts: A Survey, Journal of Artificial Intelligence Research 65 (2019) 675–782. URL: https://doi.org/10.1613/jair.1.11675.
- [9] P. Rodrigues, Processing Highly Variant Language Using Incremental Model Selection, Ph.D. thesis, Indiana University, 2012.
- [10] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, D. Varga, The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages, in: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Geona, Italy, 2006, pp. 2142–2147.
- [11] J. M. Prager, Linguini: Language Identification for Multilingual Documents, in: Proceedings

of the 32nd Annual Hawaii International Conference on Systems Sciences (HICSS-32), Maui, USA, 1999.

- [12] M. Lui, Generalized Language Identification, Ph.D. thesis, The University of Melbourne, 2014.
- [13] T. Jauhiainen, K. Lindén, H. Jauhiainen, Evaluation of language identification methods using 285 languages, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 183–191. URL: https://www.aclweb.org/anthology/W17-0221.
- [14] 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.
- [15] I. Caswell, T. Breiner, D. van Esch, A. Bapna, Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6588–6608. URL: https://www.aclweb.org/anthology/2020.coling-main.579. doi:10.18653/v1/2020. coling-main.579.
- [16] S. Góngora, N. Giossa, L. Chiruzzo, Experiments on a Guarani corpus of news and social media, in: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, Online, 2021, pp. 153–158. URL: https://aclanthology.org/2021.americasnlp-1.16. doi:10.18653/ v1/2021.americasnlp-1.16.
- [17] L. Chiruzzo, P. Amarilla, A. Ríos, G. Giménez Lugo, Development of a Guarani Spanish parallel corpus, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2629– 2633. URL: https://aclanthology.org/2020.lrec-1.320.
- [18] M. Agüero-Torales, D. Vilares, A. López-Herrera, On the logistical difficulties and findings of jopara sentiment analysis, in: Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, Association for Computational Linguistics, Online, 2021, pp. 95–102. URL: https://aclanthology.org/2021.calcs-1.12. doi:10.18653/ v1/2021.calcs-1.12.
- [19] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Gohneim, A. Hawwari, F. Al-Ghamdi, J. Hirschberg, A. Chang, P. Fung, 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, Doha, Qatar, 2014, pp. 62–72. URL: http://www.aclweb.org/anthology/W14-3907.
- [20] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, T. Solorio, Overview for the Second Shared Task on Language Identification in Code-Switched Data, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Austin, Texas, 2016, pp. 40–49. URL: https: //www.aclweb.org/anthology/W16-5805. doi:10.18653/v1/W16-5805.
- [21] K. Bar, N. Dershowitz, The Tel aviv university system for the code-switching workshop shared task, in: Proceedings of the First Workshop on Computational Approaches to Code

Switching, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 139–143. URL: https://aclanthology.org/W14-3917. doi:10.3115/v1/W14-3917.

- [22] R. Shirvani, M. Piergallini, G. S. Gautam, M. Chouikha, The Howard University system submission for the shared task in language identification in Spanish-English codeswitching, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Austin, Texas, 2016, pp. 116–120. URL: https: //aclanthology.org/W16-5815. doi:10.18653/v1/W16-5815.
- [23] G. I. Ahmad, J. Singla, Machine learning approach towards language identification of code-mixed hindi-english and urdu-english social media text, in: 2022 International Mobile and Embedded Technology Conference (MECON), 2022, pp. 215–220. doi:10.1109/ MECON53876.2022.9751958.
- [24] J. Dunn, W. Nijhof, Language identification for austronesian languages, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 6530–6539. URL: https://aclanthology. org/2022.lrec-1.701.
- [25] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, 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, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 427–431. URL: https://www.aclweb.org/anthology/ E17-2068.
- [26] A. Dutta, Word-level language identification using subword embeddings for code-mixed Bangla-English social media data, in: Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 76–82. URL: https://aclanthology.org/2022.dclrl-1.10.
- [27] L. Kevers, CoSwID, a code switching identification method suitable for under-resourced languages, in: Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, European Language Resources Association, Marseille, France, 2022, pp. 112–121. URL: https://aclanthology.org/2022.sigul-1.15.
- [28] N. Sarma, R. Sanasam Singh, D. Goswami, Switchnet: Learning to switch for word-level language identification in code-mixed social media text, Natural Language Engineering 28 (2022) 337–359. doi:10.1017/S1351324921000115.
- [29] A. Lambebo Tonja, M. Gemeda Yigezu, O. Kolesnikova, M. Shahiki Tash, G. Sidorov, A. Gelbuk, Transformer-based Model for Word Level Language Identification in Code-mixed Kannada-English Texts, arXiv e-prints (2022) arXiv:2211.14459. arXiv:2211.14459.
- [30] M. Volk, L. Fischer, P. Scheurer, B. S. Schroffenegger, R. Schwitter, P. Ströbel, B. Suter, Nunc profana tractemus. detecting code-switching in a large corpus of 16th century letters, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2901–2908. URL: https://aclanthology.org/2022.lrec-1.311.
- [31] A. F. Hidayatullah, A. Qazi, D. T. C. Lai, R. A. Apong, A systematic review on language identification of code-mixed text: Techniques, data availability, challenges, and framework development, IEEE Access 10 (2022) 122812–122831. doi:10.1109/ACCESS.2022.3223703.
- [32] T. Jauhiainen, Tekstin kielen automaattinen tunnistaminen, Master's thesis, University of

Helsinki, Helsinki, 2010.

- [33] T. Jauhiainen, Language identification in texts, Ph.D. thesis, University of Helsinki, Finland, 2019. URL: http://urn.fi/URN:ISBN:978-951-51-5131-5.
- [34] T. Jauhiainen, H. Jauhiainen, N. Partanen, K. Lindén, Uralic language identification (ULI) 2020 shared task dataset and the wanca 2017 corpora, in: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 173–185. URL: https://www.aclweb.org/anthology/2020.vardial-1.16.
- [35] C. Biemann, G. Heyer, U. Quasthoff, M. Richter, The Leipzig Corpora Collection monolingual corpora of standard size, Proceedings of Corpus Linguistic 2007 (2007).
- [36] H. Jauhiainen, T. Jauhiainen, K. Linden, Wanca in Korp: Text corpora for underresourced Uralic languages, in: J. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, K. Västi (Eds.), Proceedings of the Research data and humanities (RDHUM) 2019 conference, number 17 in Studia Humaniora Ouluensia, University of Oulu, Finland, 2019, pp. 21–40.
- [37] 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.
- [38] S. Schweter, A. Akbik, Flert: Document-level features for named entity recognition, 2020. arXiv:2011.06993.