# Italian Language and Dialect Identification and Regional French Variety Detection using Adaptive Naive Bayes

Jauhiainen, Tommi

COLING
2022-10-12

Jauhiainen , T , Jauhiainen , H & Lindén , K 2022 , Italian Language and Dialect Identification and Regional French Variety Detection using Adaptive Naive Bayes . in Y þÿ S c h e r r e r , T J a u h i a i n e n , N L j u b e a i , P N a k o v , J T i e d e m a n n & M Z a m Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects : The 29th International Conference on Computational Linguistics . International conference on computational linguistics , no. 19 , vol. 29 , COLING , Gyeongju , pp. 119-129 , International Conference on Computational Linguistics , Gyeongju , Korea, Republic of , 12/10/2022 .

http://hdl.handle.net/10138/350293

COLING

Volume 29 (2022), No. 19

# Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2022)

## The 29th International Conference on Computational Linguistics

October 12 - 17, 2022
Gyeongju, Republic of Korea

ISSN 2951-2093

# Preface

These proceedings include the 13 papers presented at the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), co-located with the 29th International Conference on Computational Linguistics (COLING). Both COLING and VarDial were held in Gyeongju, South Korea, in a hybrid format, allowing all participants to either be present on-site or join virtually.

VarDial has now reached its ninth edition and continues serving the community as the main venue for researchers interested in the computational processing of diatopic language variation. The papers accepted this year address a wide range of NLP tasks such as corpus building, part-of-speech tagging and machine translation, but also address more theoretical questions related to micro-scale variation, cognate detection, mutual intelligibility and dialectometry. We are happy to see such a diverse set of research papers advancing the state of the art of NLP for dialects, low-resource languages, and language varieties.

As in previous years, the evaluation campaign continues to be an essential part of the VarDial workshop. This year, three shared tasks were proposed: Identification of Languages and Dialects of Italy (ITDI), French Cross-Domain Dialect Identification (FDI), and Dialectal Extractive Question Answering (DialQA). All three tasks address important issues in dialect and language identification. This volume includes five system description papers prepared by the participating teams, as well as a report summarizing the results and findings of the evaluation campaign.

Finally, we would like to take this opportunity to thank the shared task organizers and the participants of the evaluation campaign for their hard work. We further thank our amazing VarDial program committee members for their thorough reviews. They have been a very important part of the workshop's success in the past years.

The VarDial workshop organizers:

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri

`http://sites.google.com/view/vardial-2022/`

**Organizers:**

Yves Scherrer - University of Helsinki (Finland)
Tommi Jauhiainen - University of Helsinki (Finland)
Nikola Ljubešić - Jožef Stefan Institute (Slovenia) and University of Zagreb (Croatia)
Preslav Nakov - Qatar Computing Research Institute, HBKU (Qatar)
Jörg Tiedemann - University of Helsinki (Finland)
Marcos Zampieri - George Mason University (USA)

**Program Committee:**

Željko Agić (Corti, Denmark)
César Aguilar (Universidad Veracruzana, Mexico)
Laura Alonso y Alemany (University of Cordoba, Argentina)
Eric Atwell (University of Leeds, United Kingdom)
Jorge Baptista (University of Algarve and INESC-ID, Portugal)
Eckhard Bick (University of Southern Denmark, Denmark)
Johannes Bjerva (University of Copenhagen, Denmark)
Francis Bond (Nanyang Technological University, Singapore)
Aoife Cahill (Educational Testing Service, United States)
David Chiang (University of Notre Dame, United States)
Paul Cook (University of New Brunswick, Canada)
Çağrı Çöltekin (University of Tübingen)
Jon Dehdari (Think Big Analytics, United States)
Liviu Dinu (University of Bucharest, Romania)
Stefanie Dipper (Ruhr University Bochum, Germany)
Sascha Diwersy (University of Montpellier, France)
Mark Dras (Macquarie University, Australia)
Tomaž Erjavec (Jožef Stefan Institute, Slovenia)
Pablo Gamallo (University of Santiago de Compostela, Spain)
Cyril Goutte (National Research Council, Canada)
Nizar Habash (New York University Abu Dhabi, UAE)
Chu-Ren Huang (Hong Kong Polytechnic University, Hong Kong)
Radu Ionescu (University of Bucharest, Romania)
Surafel Melaku Lakew (FBK , Italy)
Ekaterina Lapshinova-Koltunski (Saarland University, Germany)
Lung-Hao Lee (National Central University, Taiwan)
John Nerbonne (University of Groningen, Netherlands and University of Freiburg, Germany)
Kemal Oflazer (Carnegie-Mellon University in Qatar, Qatar)
Maciej Ogrodniczuk (IPAN, Polish Academy of Sciences, Poland)
Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
Santanu Pal (Saarland University, Germany)
Barbara Plank (LMU Munich, Germany and ITU Copenhagen, Denmark)
Taraka Rama (University of North Texas, United States)
Francisco Rangel (Autoritas Consulting, Spain)
Reinhard Rapp (University of Mainz, Germany and University of Aix-Marseille, France)
Paolo Rosso (Technical University of Valencia, Spain)
Rachel Edita O. Roxas (National University, Phillipines)

Fatiha Sadat (Université du Québec à Montréal (UQAM), Canada)
Tanja Samardžić (University of Zurich, Switzerland)
Kevin Scannell (Saint Louis University, United States)
Serge Sharoff (University of Leeds, United Kingdom)
Miikka Silfverberg (University of British Columbia, Canada)
Kiril Simov (Bulgarian Academy of Sciences, Bulgaria)
Milena Slavcheva (Bulgarian Academy of Sciences, Bulgaria)
Marco Tadić (University of Zagreb, Croatia)
Liling Tan (Rakuten Institute of Technology, Singapore)
Joel Tetreault (Dataminr, United States)
Francis Tyers (Indiana University, United States)
Pidong Wang (Google Inc., United States)
Taro Watanabe (Google Inc., Japan)

# Table of Contents

# Italian Language and Dialect Identification and Regional French Variety Detection using Adaptive Naive Bayes

**Tommi Jauhiainen, Heidi Jauhiainen, Krister Lindén**
Department of Digital Humanities, University of Helsinki, Finland
tommi.jauhiainen@helsinki.fi

## Abstract

This article describes the language identification approach used by the SUKI team in the Identification of Languages and Dialects of Italy and the French Cross-Domain Dialect Identification shared tasks organized as part of the VarDial workshop 2022. We describe some experiments and the preprocessing techniques we used for the training data in preparation for the shared task submissions, which are also discussed. Our Naive Bayes-based adaptive system reached the first position in Italian language identification and came second in the French variety identification task.

## 1 Introduction

Language identification (LI) of digital text still poses difficulties for text classification methods when performed in more complex situations (Jauhiainen et al., 2019d). One of the problematic contexts is the closeness of the languages to be identified. In this article, we tackle the problem of close language identification for languages or dialects traditionally used in Italy and distinguishing between regional French varieties used in Europe and Canada. The research and experiments were conducted while participating in the language identification shared tasks organized in connection with the ninth edition of the VarDial workshop for NLP for similar languages, varieties, and dialects (Aepli et al., 2022). The French Cross-Domain Dialect Identification (FDI) and the Identification of Languages and Dialects of Italy (ITDI) shared tasks were organized for the first time. However, they were following a long line of VarDial-related LI shared tasks from the Discriminating Between Similar languages (DSL) tasks in 2014–2017 (Zampieri et al., 2014, 2015; Malmasi et al., 2016; Zampieri et al., 2017) to newer, more specialized ones such as Romanian Dialect Identification (RDI) and Uralic Language Identification (ULI) in 2020 and 2021 (Gaman et al., 2020; Chakravarthi et al., 2021).

The ITDI shared task focused on 11 living Romance languages or dialects: Emiliano-Romagnolo (*eml*), Friulian (*fur*), Ladin (*lld*), Ligurian (*lij*), Lombard (*lmo*), Neapolitan (*nap*), Piemontese (*pms*), Sardinian (*srd*), Sicilian (*scn*), Tarantino (considered a dialect of Neapolitan by ISO 639-3), and Venetian (*vec*). The shared task was a closed one; hence, no other data besides that indicated and provided by the organizers were to be used. The organizers also stated that the test set would contain only a subset of the 11 languages.

The FDI shared task featured four regional varieties of written French from news sites in France, Belgium, Switzerland, and Canada. The organizers of the task provided all the data.

In Section 2, we present previous work on language identification of the languages that are the targets of these two shared tasks. In Section 3, we describe the data provided and allowed in the tasks, and, in Section 4, we present our method. Our experiments and the preprocessing done to the training data are explained in Section 5. In Section 6, we present and discuss the results of the submitted runs.

## 2 Previous Work

To our knowledge, there is no previous LI research focusing specifically on the languages of Italy. However, previous language identification-related research has featured the rare Romance languages that make up the ITDI repertoire. Emiliano-Romagnolo, Friulian, and Sardinian were part of the 372 languages featured in the research by Rodrigues (2012). Benedetto et al. (2002) automatically created a phylogenetic-like tree for languages based on more than 50 versions of the Universal Declaration of Human Rights, including the Friulian and the Sardinian editions. Lombard, Piemontese, Sicilian, and Venetian were featured in the ex-



*Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 119–129

periments leading to the development of the HeLI-method (Jauhiainen, 2010; Jauhiainen et al., 2016). Lombard, Neapolitan, Piemontese, Sicilian, and Venetian were included in the LI experiments conducted by Majliš (2011). King and Abney (2013) and King (2015) investigated word-level language identification in multilingual documents, including mixed Lombard - English, among other combinations. Bernier-Colborne et al. (2021) mention the Lombard - Italian pair as one of the top 10 most frequently confused pairs in the ULI-178 track of the Uralic Language Identification shared task. The ULI-178 track was a general language identification task between 178 languages, among them Lombard, Piemontese, Sardinian, Sicilian, and Venetian (Jauhiainen et al., 2020b). Neapolitan, Piemontese, and Sicilian were part of the ALTW 2010 multilingual language identification shared task dataset (Baldwin and Lui, 2010). Caswell et al. (2020) investigated language identification in the context of web crawling and mention Neapolitan, Sicilian, and Venetian as part of the lowest-resource languages in their research. All the languages of the task except Lombard were included in the language identifier featuring more than 900 languages developed by Brown (2012, 2013). Also, Lombard was added to his version with more than 1300 languages (Brown, 2014).[1] Emiliano-Romagnolo, Lombard, Neapolitan, Piemontese, Sicilian, and Venetian are part of the repertoire of the FastText off-the-shelf language identifier[2] and Lombard, Piemontese, Sardinian, and Sicilian are included in the HeLI-OTS off-the-shelf language identifier (Jauhiainen et al., 2022).[3]

Distinguishing between French regional varieties from France and Canada was part of the overall aims in the 2016 and 2017 editions of the Discriminating between Similar Languages (DSL) shared tasks (Malmasi et al., 2016; Zampieri et al., 2017).[4] The 2016 edition of the DSL was won by the *tubasfs* team using SVM and character n-grams from one to seven (Çöltekin and Rama, 2016). They managed to achieve 95.8% recall for the Canadian variety and 94.0% recall for the French variety. In 2016, we came second with a HeLI method-based identifier using words and character n-grams from one to six (Jauhiainen et al., 2016). The DSL 2017, the last one of its kind, was won by the *CECL* team using SVM with character n-grams from one to four in the first stage to detect the language group and another SVM with a variety of features in addition to character n-grams such as POS tag n-grams, the proportion of capitalized letters and punctuation marks to detect the language within the group (Bestgen, 2017).

## 3 Data

### 3.1 ITDI

In the ITDI, the participants were allowed to train their systems using the Wikipedia dumps for the 11 languages or dialects featured in the shared task. Additionally, it was possible to use the dump of the Italian language Wikipedia. All featured languages or dialects have their version of the Wikipedia online encyclopedia written in their respective language or dialect. The ISO 639-3 identifier *eml* for Emilian-Romagnol is considered deprecated as the language has been split into separate Emilian (*egl*) and Romagnol (*rgn*) languages, but Wikipedia is still shared between both languages.[5] The Sardinian, *srd*, is considered a macrolanguage in ISO 639-3, containing four separate Sardinian languages. It is possible that the Wikipedias for the *eml* and *srd* contain articles written in those separate languages. However, we did not investigate this possibility further. We did not utilize the Italian Wikipedia in any way in the experiments. The list of languages and dialects, their ISO 639-3 codes, tags used in the shared task, and the identities of their Wikipedia dumps can be seen in Table 1.

In contrast to the training data, the material for system development was provided directly by the shared task organizers. It came in one text file containing 6,799 lines which seemed to be single sentences preceded by the shared task tags. The development set included only a subset of seven of the 11 languages. The shared task participants had been informed that the test set would also be a subset of the 11 languages, but the number and identity of the missing languages were not indicated. The languages and the amount of development data for each of them can be seen in Table 2. The test set contained 11,090 lines in unknown languages or dialects.

[1] https://sourceforge.net/projects/la-strings/files/

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

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

[4] http://ttg.uni-saarland.de/resources/DSLCC/

[5] https://eml.wikipedia.org/wiki/Längua_emiglièna-rumagnôla

| Language/Dialect | ISO 639-3 | ST tag | Dump name with date | .bz2 size |
|---|---|---|---|---|
| Emiliano-Romagnolo | *eml* | **EML** | emlwiki-20220301 | 9.3 MB |
| Friulian | *fur* | **FUR** | furwiki-20220301 | 2.5 MB |
| Ladin | *lld* | **LLD** | lldwiki-20220301 | 2.8 MB |
| Ligurian | *lij* | **LIJ** | lijwiki-20220301 | 6.6 MB |
| Lombard | *lmo* | **LMO** | lmowiki-20220301 | 25 MB |
| Neapolitan | *nap* | **NAP** | napwiki-20220301 | 5.4 MB |
| Piemontese | *pms* | **PMS** | pmswiki-20220301 | 14 MB |
| Sardinian | *srd* | **SC** | scwiki-20220301 | 7.2 MB |
| Sicilian | *scn* | **SCN** | scnwiki-20220301 | 12 MB |
| Tarantino | *nap* | **ROA_TARA** | roa_tarawiki-20220301 | 6.4 MB |
| Venetian | *vec* | **VEC** | vecwiki-20220301 | 27 MB |

Table 1: The Wikipedia dumps used in the ITDI shared task.

| ST tag | lines |
|---|---|
| EML | 0 |
| FUR | 676 |
| LLD | 0 |
| LIJ | 617 |
| LMO | 1,231 |
| NAP | 0 |
| PMS | 1,191 |
| SC | 477 |
| SCN | 1,371 |
| ROA_TARA | 0 |
| VEC | 1,236 |

Table 2: The number of lines of each language in the development set of the ITDI shared task.

### 3.2 FDI

In the FDI, the participants were provided with training and development data for four regional varieties of French from France, Belgium, Switzerland, and Canada (Gaman et al., 2022). The data had been extracted from news websites in these countries using country-independent query words. A named entity recognizer (NER) Spacy had been run on the data, and all the detected entities had been changed to $\$NE\$$ in order to remove country-specific bias. The data is divided into paragraphs of three sentences or less. The amount of data for the different varieties is not balanced, as seen in Table 3. According to the data compilers, it was not easy to get Canadian material as most news sites in the country are subscription-based (Gaman et al., 2022).

In both the training and the development sets, the lines seem not to have been randomized. When we tested combining consecutive lines, they seemed to make up complete news articles or web pages. However, we expected the test set not to repeat this pattern.

## 4 Method

We used the same system we had developed for and used in the winning submission of the 2021 edition of Romanian Dialect Identification (Jauhiainen et al., 2021).[6] The system uses a Naive Bayes-based method using the observed relative frequencies of multiple-size character n-grams as probabilities. We first used the method as a baseline for the Cuneiform Language Identification (CLI) shared task (Jauhiainen et al., 2019a) and later with adaptive language models (Jauhiainen et al., 2019c) to win the Discriminating between the Mainland and Taiwan variation of Mandarin Chinese (DMT) (Jauhiainen et al., 2019b) and the RDI 2021 (Jauhiainen et al., 2021) shared tasks. The Naive Bayes type method adds together logarithms of the relative frequencies of character *n*-gram combinations $f_i$ in the training data $C_g$ as defined in Equation 1:

$$R(g,M) = -lg_{10} \prod_{i=1}^{\ell_{MF}} v_{C_g}(f_i) = \sum_{i=1}^{\ell_{MF}} -lg_{10}(v_{C_g}(f_i)) \quad (1)$$

where $\ell_{MF}$ is the number of individual features in the mystery text $M$ to be identified and $f_i$ is $M$'s $i$th feature. The relative frequency, $v_{C_g}(f)$, is calculated as in Equation 2:

$$v_{C_g}(f) = \begin{cases} \frac{c(C_g,f)}{\ell_{C_g^F}}, \textit{ if } c(C_g,f) > 0 \\ \frac{1}{\ell_{C_g^F}} pm, \textit{ otherwise} \end{cases} \quad (2)$$

[6] The implementation of the language identifier used to produce the best results for the ITDI shared task is available from GitHub at https://github.com/tosaja/TunPRF.

| Variety | Code | #lines | #tokens | Tokens per line | #NE |
|---|---|---|---|---|---|
| France | FR | 61,777 | 4,224,301 | 68 | 587,138 |
| Belgium | BE | 121,746 | 7,241,609 | 59 | 1,104,562 |
| Switzerland | CH | 141,261 | 8,494,657 | 60 | 1,112,525 |
| Canada | CA | 34,003 | 1,694,760 | 50 | 184,083 |

Table 3: The training and development datasets sizes for the FDI shared task.

where $c(C_g, f)$ is the count of feature $f$ in the training corpus $C_g$ of the language $g$. $\ell_{C_g^F}$ is the length of the corpus $C_g$ when it has been transformed into a collection of features $F$, e.g., features of the same type as $f$. The $pm$ is the penalty modifier, which is optimized using the development data.

The system uses an adaptation technique to learn from the test data (Jauhiainen et al., 2019c). There is also a possibility to perform iterative adaptation, in which the test data is processed several times from the beginning of the adaptation process.

The exact range of the used character n-grams is optimized using the development data. After optimizing the basic method, the parameters for the adaptive version are determined. In the adaptation technique, the test data is first identified with the basic method, and a confidence score is calculated for each identified instance. The confidence score is the difference between the scores of the best and the second-best language. The test instances are sorted according to the confidence scores and then divided into a certain number of splits. The number of splits is determined using the development data. The character n-gram frequency information from the most confident split is added to the respective language models, and the rest of the material is re-identified with the adapted models. Then the rest of the material is again sorted and divided into equally sized splits, and the information from the most confident split is added to the models and the rest re-identified. The previous process is repeated until all the material is added to the language models.

In the iterative version, the adaptation process is restarted from the beginning. The number of possible iterations is also determined using the development set.

## 5 Experiments

This section presents the details of the experiments and various preprocessing techniques we used when participating in the shared tasks.

### 5.1 ITDI

The organizers of the Identification of Languages and Dialects of Italy shared task provided a script that could be used to generate a .json file from the .bz2 files downloaded from Wikipedia. Instead of using the script, we created plain text versions of the dumps using the command:

```
-m wikiextractor.WikiExtractor
xxxwiki-20220301- ... .bz2 -o xxx_texts
```

The training data extracted this way contained 1.91 million lines, some extended text passages, and some just short headings, names, or even empty lines. The first thing we did was to remove duplicate lines within each language or dialect. This deduplication procedure reduced the size of the training data to 0.93 million lines.

As the next step, we removed the lines containing wiki markup, which we found by using the following regular expressions:

```
&lt;comment&gt;.*&lt;/comment&gt;
&lt;contributor&gt;
&lt;/contributor&gt;
&lt;format&gt;.*&lt;/format&gt;
&lt;ip&gt;.*&lt;/ip&gt;
&lt;minor /&gt;
&lt;model&gt;.*&lt;/model&gt;
&lt;ns.*/ns&gt;
&lt;parentid&gt;.*&lt;/parentid&gt;
&lt;revision&gt;
&lt;timestamp&gt;.*&lt;/timestamp&gt;
&lt;username&gt;.*&lt;/username&gt;
```

We also removed all lines with a tab character followed by a lowercase letter and unified the numbers so that all number characters were changed to "1". Then, we again removed duplicate lines which left us with 880,000 lines. At this point, we took an inventory of the number of lines for each language and dialect (Table 4).

At each stage, we had tested the performance of our Naive Bayes-based system on the development data. At this stage, the Lombard language had the worst precision with 83.2%, and we decided to try and clean its training data to improve its precision. As it seemed that, in general, shorter lines were of lower quality than longer ones, we removed all Lombard lines with less than 14 characters from

| ST tag | # lines |
|---|---|
| EML | 16,425 |
| FUR | 18,040 |
| LLD | 31,524 |
| LIJ | 38,645 |
| LMO | 185,116 |
| NAP | 34,327 |
| PMS | 208,485 |
| SC | 41,169 |
| SCN | 95,693 |
| ROA_TARA | 36,818 |
| VEC | 173,452 |

Table 4: The number of lines for each language or dialect in the ITDI training data after unifying the number characters.

the training corpus. We also did further cleaning of the wiki markup for all languages by removing lines using the following regular expressions:

```
&lt;/math&gt;
&lt;/[pP]oem&gt;
&lt;/small&gt;
&lt;references&gt;
&lt;/html&gt;
&lt;/includeonly&gt;&lt;/onlyinclude&gt;
&lt;/table&gt;
&lt;?php
&lt;BR C.* &gt;
#redirect
```

We removed all lines that did not include lowercase ASCII characters and the remaining "<" and "br>" tags. Once more, we removed any possible duplicate lines. Our efforts to improve the precision of LMO were in vain, as it dropped from 83.2% to 83.0%. However, the micro F1 over all the languages remained the same, 92.7, so we kept the changes.

At this stage, the Venetian language had the worst recall of all the languages, 75.6%. While taking a look at the erroneously identified sentences, we noticed that, in fact, part of the Venetian Wikipedia used a slightly different orthography than the development data. The Wikipedia dumps contained the "ł" and "Ł" characters, whereas only "l" and "L" were present in the development data. We used a simple regular expression to change the training data to correspond with the development data. This unification of orthographies improved the recall of Venetian from 75.6% to 79.1% and the precision of Lombard from 83.0% to 85.6%. The overall micro F1 also increased slightly from 92.7 to 93.3.

One phenomenon we were aware of due to our previous experiences using Wikipedia dumps as training material was that some of the smaller Wikipedias might contain relatively large parts automatically generated from a database. In particular, the pages describing the French municipalities are usually generated using templates. These template-based articles were also found as part of the Venetian Wikipedia: out of the 173,091 Venetian lines, 33,701 were automatically generated information about French communes. We removed the lines using the following regular expression to detect them:

```
egrep -v 'el xe on comun de.*abitanti del
departemento.*in Fransa\.'
```

The recall of Venetian increased from 79,1% to 84,0%, and at the same time, the precision of Lombard rose from 85.6% to 88.2%.

Venetian still had the lowest individual F1 score, 89.9. We aimed to increase further the quality of the training set by first removing all lines which did not have a word beginning with a lowercase ASCII letter and then removing all the 2,983 lines explaining roman numbers such as:

```
El 11 (LXIII en numeri romani) el xe ...
El 11 (LXIV en numeri romani) el xe ...
El 11 (LXIX in numeri romani) el xe ...
El 11 (LXV en numeri romani) el xe ...
```

Additionally, in a similar manner to the French towns, we removed the municipalities of Italy listed in the Venetian Wikipedia. These cleaning operations resulted in a slight increase of the F1 to 90.3. The overall micro F1 had by now risen from 93.3 to 94.2.

Further cleaning, e.g., removing lines describing the Spanish towns in the Venetian Wikipedia and removing all lines containing specific additional wiki markup, did not improve the overall F1 score. As further preprocessing seemed less fruitful, we started experimenting with the adaptive version of the Naive Bayes identifier, with which evaluating new versions of the training corpora would take much longer.

We tested the adaptive version with 128, 256, and 512 splits. Additionally, we tested with 128 splits and two iterations. They all returned the same micro F1 of 96.2, which was higher than the score of 94.2, which was attained without adaptation.[7]

[7] If time allowed, one would begin finding the optimal number of splits from two and then double the number of splits every iteration as we did with FDI (see Table 6). Due to time constraints, for ITDI, we skipped the first ones. If 128 splits had produced better results than 256 splits, we would

In the end, we opted to continue using 512 splits with only one adaptation round.

Afterward, we still decided to continue training corpora cleaning and removed additional template-generated text from Lombard training data. Additionally, we removed from all languages those lines that did not include a space character followed by a lowercase ASCII alphabetical character, e.g., those that did not have a word starting with a lowercase letter. These modifications did not improve the results, but we still decided to use them as we considered the training data to be in better shape.

As stated by the organizers and indicated by the languages missing from the development set, the test data would not include all the languages in the training data. Furthermore, we were unsure what measure would be used to evaluate the submissions. These facts led us to prepare one submission in preparation for the measure being macro F1. Also, we hoped that leaving out unnecessary languages might help to boost the performance of the remaining languages. We have previously developed a method for language set identification (Jauhiainen et al., 2015) and used it while collecting rare Uralic languages from the internet (Jauhiainen et al., 2020a). However, instead of using our language set identification method, we devised a simple thresholding method to leave out the most probable unnecessary languages using the development set as a guideline. Based on the development set, we surmised that it would be safe to remove from the repertoire those languages that, after all lines had been identified once, had been assigned fewer lines than 10% of the average number of lines for each language.

### 5.2 FDI

In the French Cross-Domain Dialect Identification shared task, the training data seemed to be of better quality than in the ITDI, and when perusing it, we did not notice any need for extensive preprocessing. We started with optimizing the parameters for the Naive Bayes identifier. Our first optimization run gave the best result, a micro F1 of 0.646, with just character six-grams, which were the maximum size for n-grams on that run.

We noticed that the training data for some language varieties contained a large amount of repetition, as seen in Table 5. Especially the Canadian variety training corpus consisted of identical lines repeated hundreds of times. Even the rarest lines in that corpus are repeated 55 times.

| Code | #lines | #unique lines |
|---|---|---|
| FR | 61,777 | 55,927 |
| BE | 121,746 | 113,487 |
| CH | 141,261 | 107,982 |
| CA | 34,003 | 169 |

Table 5: The number of lines in the FDI training data before and after removing duplicates.

We did the same initial optimization run with the deduplicated training data and ended up with a micro F1 of 0.634. The score was slightly worse than before deduplication, so we continued experimenting with the original training set. We also tested lowercasing the training data and the mystery texts, which gave lower micro F1 scores with both original and deduplicated datasets. Further optimization with higher order n-grams led us to use only character eight-grams and the penalty modifier of 1.26, which gave a micro F1 score of 0.675 on the development data. We then optimized with the deduplicated training data, which resulted in eight grams and a penalty modifier of 1.73, giving a micro F1 of 0.659, which was again lower than without deduplication. Next, we experimented with removing the named entity tags from the training and the development data, which again resulted in a slightly lower micro F1 of 0.659.

So far, we had used the micro F1 as our guideline when optimizing the system even though we were aware that the macro F1 would be used to rank the official submissions. The reason for using micro F1 was that our optimization system did not produce correct macro F1 scores, which we fixed at this stage. The macro F1 corresponding to micro F1 of 0.675 was 0.495. We then proceeded to experiment with the adaptive version of the identifier.

We evaluated several combinations of the number of splits and iteration rounds as seen in Table 6. In the end, we used the character eight-grams, the penalty modifier of 1.26, 128 splits, and three iterations, giving us a macro F1 of 0.553 and a micro F1 of 0.745 on the development set.

As a last experiment, we decided to try to unify the numbers in a similar way we did with the ITDI data using the non-adaptive version of the system. Unifying the numbers increased the macro F1 from 0.495 to 0.498 and the micro F1 from 0.675 to 0.681. Due to limited time, we could not run the

have experimented with 64 splits and continued reducing the number of splits as long as the results improved.

| # splits | # iterations | Macro F1 | Micro F1 |
|---|---|---|---|
| 2 | 4 | 0.510 | 0.698 |
| 4 | 4 | 0.523 | 0.713 |
| 8 | 4 | 0.537 | 0.729 |
| 16 | 4 | 0.550 | 0.742 |
| 32 | 3 | 0.552 | 0.744 |
| 64 | 3-4 | 0.5526 | 0.7450 |
| 128 | 3-4 | **0.5529** | **0.7454** |
| 256 | 3-4 | 0.5527 | 0.7451 |
| 512 | 3-4 | 0.5525 | 0.7449 |
| 1024 | 3-4 | 0.5524 | 0.7447 |

Table 6: Optimizing adaptation parameters with the FDI training and development data. The best scores are bolded.

adaptive version on development data. As unifying improved the results, we decided to unify the numbers with the adaptive version for the actual submissions.

## 6 Results

In this section, we describe the results of the submitted runs and the conclusions we can derive from them.

### 6.1 ITDI

We submitted three sets of predictions for the ITDI task. The main difference between the submissions was the data used to train the identifier. All the submissions used character n-grams from three to eight with a penalty modifier of 2.1. The number of splits in adaptation was set to 512, and iterative adaptation was not used. We added a space character to the beginning and the end of the text to be identified so that our identifier would recognize the beginning of the first word and the end of the last word.

The first submission used combined training and development data, and the second just the training data. The third system combined the training and development data but without the data for Piemontese and Sardinian, which were discarded in the language set identification phase due to having less than the threshold amount of instances. The weighted average F1 score for all of our three submissions was quite similar and on a completely different level from the results of the best submissions of the two other participating teams, as seen in Table 7. The best baseline provided by the organizers was closer to our results than those of the other teams but still substantially behind them.

| Team | Submission | Wgt. F1 | Mac. F1 |
|---|---|---|---|
| SUKI | 2 | 0.9007 | 0.6729 |
| SUKI | 1 | 0.8983 | 0.6714 |
| SUKI | 3 | 0.8982 | 0.7458 |
| Org. | Baseline | 0.7726 | 0.5193 |
| Phlyers | 3 | 0.6943 | 0.5379 |
| ETHZ | 3 | 0.6880 | 0.4828 |

Table 7: The results of the ITDI shared task with the best baseline.

Without further experiments, it is difficult to say whether the distance to the other teams is due to differences in preprocessing Wikipedia or to using adaptive language models. The results on the language/dialect level can be seen in Table 8. The worst performing language of the languages present in the test set was the Neapolitan dialect Tarantino.

### 6.2 FDI

All our submissions to the FDI shared task used character eight grams with a penalty modifier of 1.26. The number of splits in adaptation was set to 128, and three iterations of adaptation to the test data were used. As in the ITDI task, the difference between the submissions comes from the data used for training. The first submission uses the training data, the second uses the development data, and the third uses a combination of both.

In contrast to the ITDI, the FDI shared task was ranked using the macro F1 score. The macro F1 score of our best submission, 0.266, was far behind the 0.344 of the winning NRC team. However, the results of the NRC team were still clearly lower than that of the best baseline, CamemBERT (Gaman et al., 2022), as seen in Table 9.

Table 10 is a confusion matrix for our third and best run. The contrast to our second run in Table 11 is dramatic. In the third run, only 28 lines were identified as the Canadian variety as opposed to the 15,518 lines identified as the Swiss variety. In the second run, 6,188 lines were identified as the Canadian variety and only 28 as the Swiss variety. It seems that the choice of training data is mostly responsible for these great differences.

The difference between our results on the development data vs. the test data is quite significant compared to similar results reported by the organizers (Gaman et al., 2022). Table 12 shows how our Macro F1 drops over 50% from development

| Language | Precision | Recall | F1 score | Lines in test |
|---|---|---|---|---|
| EML | 0.8916 | 0.9273 | 0.9091 | 825 |
| FUR | 0.9969 | 0.9781 | 0.9874 | 1,323 |
| LIJ | 0.9831 | 0.9947 | 0.9889 | 2,282 |
| LLD | 0.9971 | 0.9268 | 0.9607 | 2,200 |
| LMO | 0.8991 | 0.9826 | 0.9390 | 689 |
| NAP | 0.8927 | 0.887 | 0.8898 | 2,026 |
| ROA_TARA | 0.7532 | 0.0962 | 0.1706 | 603 |
| SC | 0 | 0 | 0 | 0 |
| SCN | 0 | 0 | 0 | 0 |
| VEC | 0.7929 | 0.9982 | 0.8838 | 1,139 |

Table 8: Per language results for our best submission on the ITDI shared task.

| Team | Submission | Macro F1 | Weighted F1 | Micro F1 |
|---|---|---|---|---|
| CamemBERT | baseline | 0.3967 | - | 0.5584 |
| NRC | 2 | 0.3437 | 0.4581 | 0.4936 |
| SUKI | 3 | 0.2661 | 0.3422 | 0.3918 |
| DontClassify | 1 | 0.2627 | 0.3236 | 0.3914 |
| SUKI | 1 | 0.2603 | 0.3439 | 0.3984 |
| SUKI | 2 | 0.1383 | 0.1958 | 0.2339 |

Table 9: The results of the FDI shared task.

| | BE | CA | CH | FR | Recall | Precision | F1 score |
|---|---|---|---|---|---|---|---|
| **BE** | 7,252 | 1 | 7,119 | 863 | 47.6% | 39.4% | 43.1 |
| **CA** | 97 | 16 | 574 | 257 | 1.7% | 57.1% | 3.3 |
| **CH** | 2,148 | 1 | 6,570 | 1,105 | 66.9% | 42.3% | 51.9 |
| **FR** | 8,912 | 10 | 1,255 | 553 | 5.2% | 19.9% | 8.2 |

Table 10: The confusion matrix for our third and best submission at the FDI shared task, with the recall, precision, and the F1 score for each variety.

| | BE | CA | CH | FR | Recall | Precision | F1 score |
|---|---|---|---|---|---|---|---|
| **BE** | 7,262 | 5,220 | 5 | 2,748 | 47.7% | 31.3% | 37.8 |
| **CA** | 717 | 162 | 2 | 63 | 17.2% | 2.6% | 4.5 |
| **CH** | 5,940 | 568 | 7 | 3,309 | 0.1% | 25.0% | 0.1 |
| **FR** | 9,317 | 238 | 14 | 1,161 | 10.8% | 15.9% | 12.9 |

Table 11: The confusion matrix for our second submission at the FDI shared task, with the recall, precision, and the F1 score for each variety.

to testing, whereas the drop for the best baseline is less than 20%. After the shared tasks, we created a version of our Naive Bayes system, which automatically determines the best parameters using the development set. Using the new implementation, we conducted some further experiments with the language annotated test data, and now it is clear that the optimal character n-gram range for the test data differs significantly from that of the development data. The optimal character range for the test data seems to be from four to seven characters, whereas it is from eight to eight for the development data. With the optimal character n-gram range, the NB identifier gets a macro F1 score of 0.3539 without language model adaptation. This score would be more comparable with the scores of the winning NRC submission. However, the real issue was the language model adaptation which lowered the results considerably. On the one hand, even using just the character eight-grams without adaptation gives the macro F1 score of 0.3306 on the test data, and on the other hand, using character n-grams from four to seven, the optimal range, with adaptation results in macro F1 score of 0.2628.

## Acknowledgements

The research was conducted within the Language Identification of Speech and Text project funded by the Finnish Research Impact Foundation from its Tandem Industry Academia funding in cooperation with Lingsoft.

## References

Noëmi Aepli, Antonios Anastasopoulos, Adrian Chifu, William Domingues, Fahim Faisal, Mihaela Găman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea.

Timothy Baldwin and Marco Lui. 2010. Multilingual language identification: ALTW 2010 shared task data. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 4–7, Melbourne, Australia.

Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language Trees and Zipping. *Physical Review Letters*, 88(4).

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2021. N-gram and Neural Models for Uralic Language Identification: NRC at VarDial 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 128–134, Kyiv, Ukraine. Association for Computational Linguistics.

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain. Association for Computational Linguistics.

Ralf Brown. 2014. Non-linear Mapping for Improved Identification of 1300+ Languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar. Association for Computational Linguistics.

Ralf D. Brown. 2012. Finding and Identifying Text in 900+ Languages. *Digital Investigation*, 9:S34–S43.

Ralf D. Brown. 2013. Selecting and Weighting N-grams to Identify 1100 Languages. In *Proceedings of the 16th International Conference on Text, Speech and Dialogue (TSD 2013)*, pages 475–483, Plzeň, Czech Republic.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. 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*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bharathi Raja Chakravarthi, Mihaela Gaman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kyiv, Ukraine. Association for Computational Linguistics.

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

Mihaela Gaman, Adrian Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2022. FreCDo: A New Corpus for Large-Scale French Cross-Domain Dialect Identification. *(under review)*.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A Report on the VarDial Evaluation Campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14,

| System | Source data | Target data | Macro F1 | Micro F1 |
|---|---|---|---|---|
| CamemBERT | Training | Development | 0.4784 | 0.7352 |
| CamemBERT | Training | Testing | 0.3967 | 0.5584 |
| Adaptive NB | Training | Development | 0.5529 | 0.7454 |
| Adaptive NB | Training + Development | Testing | 0.2661 | 0.3918 |
| Adaptive NB | Training | Testing | 0.2603 | 0.3984 |

Table 12: Comparison between the baseline results and our best submission.


Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2020a. Building Web Corpora for Minority Languages. In *Proceedings of the 12th Web as Corpus Workshop*, pages 23–32, Marseille, France. European Language Resources Association.

Tommi Jauhiainen. 2010. Tekstin kielen automaattinen tunnistaminen. Master's thesis, University of Helsinki, Helsinki.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019a. Language and Dialect Identification of Cuneiform Texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 89–98, Ann Arbor, Michigan. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2019b. Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2021. Naive Bayes-based experiments in Romanian dialect identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 76–83, Kyiv, Ukraine. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020b. 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*, pages 173–185, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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. 2019c. 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. 2019d. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Ben King and Steven Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, USA.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages.* Ph.D. thesis.

Martin Majliš. 2011. Large Multilingual Corpus. Master's thesis, Charles University in Prague, Prague.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Paul Rodrigues. 2012. *Processing Highly Variant Language Using Incremental Model Selection*. Ph.D. thesis, Indiana University.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings*

*of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.