Automatic Language Identification in Texts: A Survey

Article *in* Journal of Artificial Intelligence Research · April 2018

DOI: 10.1613/jair.1.11675

CITATIONS
41

READS
2,321

5 authors, including:

Tommi Jauhiainen
University of Helsinki
18 PUBLICATIONS **104** CITATIONS
SEE PROFILE

Marcos Zampieri
Rochester Institute of Technology
107 PUBLICATIONS **2,244** CITATIONS
SEE PROFILE

Timothy Baldwin
University of Melbourne
314 PUBLICATIONS **8,040** CITATIONS
SEE PROFILE

Krister Lindén
University of Helsinki
118 PUBLICATIONS **737** CITATIONS
SEE PROFILE

Journal of Artificial Intelligence Research () Submitted 10/2018; published - Under Review

Automatic Language Identification in Texts: A Survey

Tommi Jauhiainen TOMMI.JAUHIAINEN@HELSINKI.FI
Department of Digital Humanities
The University of Helsinki

Marco Lui SAFFSD@GMAIL.COM
School of Computing and Information Systems
The University of Melbourne

Marcos Zampieri M.ZAMPIERI@WLV.AC.UK
Research Institute in Information and Language Processing
University of Wolverhampton

Timothy Baldwin TB@LDWIN.NET
School of Computing and Information Systems
The University of Melbourne

Krister Lindén KRISTER.LINDEN@HELSINKI.FI
Department of Digital Humanities
The University of Helsinki

Abstract

Language identification ("LI") is the problem of determining the natural language that a document or part thereof is written in. Automatic LI has been extensively researched for over fifty years. Today, LI is a key part of many text processing pipelines, as text processing techniques generally assume that the language of the input text is known. Research in this area has recently been especially active. This article provides a brief history of LI research, and an extensive survey of the features and methods used in the LI literature. We describe the features and methods using a unified notation, to make the relationships between methods clearer. We discuss evaluation methods, applications of LI, as well as *off-the-shelf* LI systems that do not require training by the end user. Finally, we identify open issues, survey the work to date on each issue, and propose future directions for research in LI.

1. Introduction

Language identification ("LI") is the task of determining the natural language that a document or part thereof is written in. Recognizing text in a specific language comes naturally to a human reader familiar with the language. Table 1 presents excerpts from Wikipedia articles in different languages on the topic of Natural Language Processing ("NLP"), labeled according to the language they are written in. Without referring to the labels, readers of this article will certainly have recognized at least one language in Table 1, and many are likely to be able to identify all the languages therein.

English	Natural language processing is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.
Italian	L'Elaborazione del linguaggio naturale è il processo di trattamento automatico mediante un calcolatore elettronico delle informazioni scritte o parlate nel linguaggio umano o naturale.
Chinese	自然語言處理是人工智慧和語言學領域的分支學科。
Japanese	自然言語処理は、人間が日常的に使っている自然言語をコンピュータに処理させる一連の技術であり、人工知能と言語学の一分野である。

Table 1: Excerpts from Wikipedia articles on NLP in different languages.

Research into LI aims to mimic this human ability to recognize specific languages. Over the years, a number of computational approaches have been developed that, through the use of specially-designed algorithms and indexing structures, are able to infer the language being used without the need for human intervention. The capability of such systems could be described as super-human: an average person may be able to identify a handful of languages, and a trained linguist or translator may be familiar with many dozens, but most of us will have, at some point, encountered written texts in languages they cannot place. However, LI research aims to develop systems that are able to identify *any* human language, a set which numbers in the thousands (Simons and Fennig, 2017).

In a broad sense, LI applies to any modality of language, including speech, sign language, and handwritten text, and is relevant for all means of information storage that involve language, digital or otherwise. However, in this survey we limit the scope of our discussion to LI of written text stored in a digitally-encoded form.

Research to date on LI has traditionally focused on *monolingual* documents (Hughes et al., 2006) (we discuss LI for multilingual documents in Section 10.6). In monolingual LI, the task is to assign each document a unique language label. Some work has reported near-perfect accuracy for LI of large documents in a small number of languages, prompting some researchers to label it a "solved task" (McNamee, 2005). However, in order to attain such accuracy, simplifying assumptions have to be made, such as the aforementioned monolinguality of each document, as well as assumptions about the type and quantity of data, and the number of languages considered.

The ability to accurately detect the language that a document is written in is an enabling technology that increases accessibility of data and has a wide variety of applications. For example, presenting information in a user's native language has been found to be a critical factor in attracting website visitors (Kralisch and Mandl, 2006). Text processing techniques developed in natural language processing and Information Retrieval ("IR") generally presuppose that the language of the input text is known, and many techniques assume that all documents are in the same language. In order to apply text processing techniques to real-world data, automatic LI is used to ensure that only documents in relevant languages are subjected to further processing. In information storage and retrieval, it is common to index documents in a multilingual collection by the language that they are written in, and

LI is necessary for document collections where the languages of documents are not known a-priori, such as for data crawled from the World Wide Web. Another application of LI that predates computational methods is the detection of the language of a document for routing to a suitable translator. This application has become even more prominent due to the advent of Machine Translation ("MT") methods: in order for MT to be applied to translate a document to a target language, it is generally necessary to determine the source language of the document, and this is the task of LI. LI also plays a part in providing support for the documentation and use of low-resource languages. One area where LI is frequently used in this regard is in linguistic corpus creation, where LI is used to process targeted web crawls to collect text resources for low-resource languages.

A large part of the motivation for this article is the observation that LI lacks a "home discipline", and as such, the literature is fragmented across a number of fields, including NLP, IR, machine learning, data mining, social medial analysis, computer science education, and systems science. This has hampered the field, in that there have been many instances of research being carried out with only partial knowledge of other work on the topic, and the myriad of published systems and datasets.

Finally, it should be noted that this survey does not make a distinction between languages, language varieties, and dialects. Whatever demarcation is made between languages, varieties and dialects, a LI system is trained to identify the associated document classes. Of course, the more similar two classes are, the more challenging it is for a LI system to discriminate between them. Training a system to discriminate between similar languages such as Croatian and Serbian (Ljubešić and Kranjcić, 2014), language varieties like Brazilian and European Portuguese (Zampieri and Gebre, 2012), or a set of Arabic dialects (Zampieri et al., 2015b) is more challenging than training systems to discriminate between, for example, Japanese and Finnish. Even so, as evidenced in this article, from a computational perspective, the algorithms and features used to discriminate between languages, language varieties, and dialects are identical.

2. LI as Text Categorization

LI is in some ways a special case of text categorization, and previous research has examined applying standard text categorization methods to LI (Cavnar and Trenkle, 1994; Elworthy, 1998).

(Sebastiani, 2002, Section 2.1) provides a definition of text categorization, which can be summarized as the task of mapping a document onto a pre-determined set of classes. This is a very broad definition, and indeed one that is applicable to a wide variety of tasks, amongst which falls modern-day LI. The archetypal text categorization task is perhaps the classification of newswire articles according to the topics that they discuss, exemplified by the Reuters-21578 dataset (Debole and Sebastiani, 2005). However, LI has particular characteristics that make it different from typical text categorization tasks:

1. Text categorization tends to use statistics about the frequency of words to model documents, but for LI purposes there is no universal notion of a *word*: LI must cater for languages where whitespace is not used to denote word boundaries. Furthermore, the determination of the appropriate word tokenization strategy for a given document

presupposes knowledge of the language the document is written in, which is exactly what we assume we *don't* have access to in LI.

2. In text categorization tasks, the set of labels usually only applies to a particular dataset. For example, it is not meaningful to ask which of the Reuters-21578 labels is applicable to the abstract of a biomedical journal article. However, in LI there is a clear notion of language that is independent of domain: it is possible to recognize that a text is in English regardless of whether it is from a biomedical journal, a microblog post, or a newspaper article.

3. In LI, classes can be somewhat multi-modal, in that text in the same language can sometimes be written with different orthographies and stored in different encodings, but correspond to the same class.

4. In LI, labels are non-overlapping and mutually exclusive, meaning that a text can only be written in one language. This does not preclude the existence of multilingual documents which contain text in more than one language, but when this is the case, the document can always be uniquely divided into monolingual segments. This is in contrast to text categorization involving multi-labeled documents, where it is generally not possible to associate specific segments of the document with specific labels.

These distinguishing characteristics present unique challenges and offer particular opportunities, so much so that research in LI has generally proceeded independently of text categorization research. In this survey, we will examine the common themes and ideas that underpin research in LI. We begin with a brief history of research that has led to modern LI (Section 3), and then proceed to review the literature, first introducing the mathematical notation used in the article (Section 4), and then providing synthesis and analysis of existing research, focusing specifically on the representation of text (Section 5) and the learning algorithms used (Section 6). We examine the methods for evaluating the quality of the systems (Section 7) as well as the areas where LI has been applied (Section 8), and then provide an overview of "off-the-shelf" LI systems (Section 9). We conclude the survey with a discussion of the open issues in LI (Section 10), enumerating issues and existing efforts to address them, as well as charting the main directions where further research in LI is required.

2.1 Previous Surveys

Although there are some dedicated survey articles, these tend to be relatively short; there have not been any comprehensive surveys of research in automated LI of text to date. The largest survey so far can be found in the literature review of Marco Lui's PhD thesis (Lui, 2014), which served as an early draft and starting point for the current article. Zampieri (2016) provides a historical overview of language identification focusing on the use of n-gram language models. Qafmolla (2017) gives a brief overview of some of the methods used for LI, and Garg et al. (2014) provide a review of some of the techniques and applications used previously. Shashirekha (2014) gives a short overview of some of the challenges, algorithms and available tools for LI. Juola (2006) provides a brief summary of LI, how it relates to other research areas, and some outstanding challenges, but only does so in general terms

and does not go into any detail about existing work in the area. Another brief article about LI is Muthusamy and Spitz (1997), which covers LI both of spoken language as well as of written documents, and also discusses LI of documents stored as images rather than digitally-encoded text.

3. A Brief History of LI

LI as a task predates computational methods – the earliest interest in the area was motivated by the needs of translators, and simple manual methods were developed to quickly identify documents in specific languages. The earliest known work to describe a functional LI program for text is by Mustonen (1965), a statistician, who used multiple discriminant analysis to teach a computer how to distinguish, at the word level, between English, Swedish and Finnish. Mustonen compiled a list of linguistically-motivated character-based features, and trained his language identifier on 300 words for each of the three target languages. The training procedure created two discriminant functions, which were tested with 100 words for each language. The experiment resulted in 76% of the words being correctly classified; even by current standards this percentage would be seen as acceptable given the small amount of training material, although the composition of training and test data is not clear, making the experiment unreproducible.

In the early 1970s, Nakamura (1971) considered the problem of automatic LI. According to Rau (1974) and the available abstract of Nakamura's article,[1] his language identifier was able to distinguish between 25 languages written with the Latin alphabet. As features, the method used the occurrence rates of characters and words in each language. From the abstract it seems that, in addition to the frequencies, he used some binary presence/absence features of particular characters or words, based on manual LI.

Rau (1974) wrote his master's thesis "Language Identification by Statistical Analysis" for the Naval Postgraduate School at Monterey, California. The continued interest and the need to use LI of text in military intelligence settings is evidenced by the recent articles of, for example, Rafidha Rehiman et al. (2013), Rowe et al. (2013), Tratz (2014), and Voss et al. (2014). As features for LI, Rau (1974) used, e.g., the relative frequencies of characters and character bigrams. With a majority vote classifier ensemble of seven classifiers using Kolmogor-Smirnov's Test of Goodness of Fit and Yule's characteristic (K), he managed to achieve 89% accuracy over 53 characters when distinguishing between English and Spanish. His thesis actually includes the identifier program code (for the IBM System/360 Model 67 mainframe), and even the language models in printed form.

Much of the earliest work on automatic LI was focused on identification of spoken language, or did not make a distinction between written and spoken language. For example, the work of House and Neuburg (1977) is primarily focused on LI of spoken utterances, but makes a broader contribution in demonstrating the feasibility of LI on the basis of a statistical model of broad phonetic information. However, their experiments do not use actual speech data, but rather "synthetic" data in the form of phonetic transcriptions derived from written text.

1. We were unable to obtain the original article, so our account of the paper is based on the abstract and reports in later published articles.

Another subfield of speech technology, speech synthesis, has also generated a considerable amount of research in the LI of text, starting from the 1980s. In speech synthesis, the need to know the source language of individual words is crucial in determining how they should be pronounced. Church (1985) uses the relative frequencies of character trigrams as probabilities and determines the language of words using a Bayesian model. Church explains the method – that has since been widely used in LI – as a small part of an article concentrating on many aspects of letter stress assignment in speech synthesis, which is probably why Beesley (1988) is usually attributed to being the one to have introduced the aforementioned method to LI of text. As Beesley's article concentrated solely on the problem of LI, this single focus probably enabled his research to have greater visibility. The role of the program implementing his method was to route documents to MT systems, and Beesley's paper more clearly describes what has later come to be known as a character n-gram model. The fact that the distribution of characters is relatively consistent for a given language was already well known.

The highest-cited early work on automatic LI is Cavnar and Trenkle (1994). Cavnar and Trenkle's method (which we describe in detail in Section 6.6) builds up per-document and per-language profiles, and classifies a document according to which language profile it is most similar to, using a rank-order similarity metric. They evaluate their system on 3478 documents in eight languages obtained from USENET newsgroups, reporting a best overall LI accuracy of 99.8%. Gertjan van Noord produced an implementation of the method of Cavnar and Trenkle named `TextCat`, which has become eponymous with the method itself. `TextCat` is packaged with pre-trained models for a number of languages, and so it is likely that the strong results reported by Cavnar and Trenkle, combined with the ready availability of an "off-the-shelf" implementation, has resulted in the exceptional popularity of this particular method. Cavnar and Trenkle (1994) can be considered a milestone in automatic LI, as it popularized the use of automatic methods on character n-gram models for LI, and to date the method is still considered a benchmark for automatic LI.

4. On Notation

This section introduces the notation used throughout this article to describe LI methods. We have translated the notation in the original papers to our notation, to make it easier to see the similarities and differences between the LI methods presented in the literature. The formulas presented could be used to implement language identifiers and re-evaluate the studies they were originally presented in.

A corpus C consists of individual tokens u which may be bytes, characters or words. C is comprised of a finite sequence of individual tokens, $u_1, ..., u_{l_C}$. The total count of individual tokens u in C is denoted by l_C. In a corpus C with non-overlapping segments S, each segment is referred to as C_s, which may be a short document or a word or some other way of segmenting the corpus. The number of segments is denoted as l_S.

A feature f is some countable characteristic of the corpus C. When referring to the set of all features F in a corpus C, we use C^F, and the number of features is denoted by l_{C^F}. A set of unique features in a corpus C is denoted by $U(C)$. The number of unique features is referred to as $|U(C)|$. The count of a feature f in the corpus C is referred to as $c(C, f)$. If a corpus is divided into segments S, the count of a feature f in C is defined as

the sum of counts over the segments of the corpus, i.e. $c(C, f) = \sum_{s=1}^{l_S} c(C_s, f)$. Note that the segmentation may affect the count of a feature in C as features do not cross segment borders.

A frequently-used feature is an n-gram, which consists of a contiguous sequence of n individual tokens. An n-gram starting at position i in a corpus segment is denoted $u_{i,...,i+n-1}$, where positions $i+1, ..., i+n-1$ remain within the same segment of the corpus as i. If $n = 1$, f is an individual token. When referring to all n-grams of length n in a corpus C, we use C^n and the count of all such n-grams is denoted by l_{C^n}. The count of an n-gram f in a corpus segment C_s is referred to as $c(C_s, f)$ and is defined by Equation 1:

$$c(C_s, f) = \sum_{i=1}^{l_{C_s}+1-n} \begin{cases} 1 & \text{, if } f = u_{i,...,i-1+n} \\ 0 & \text{, otherwise} \end{cases} \quad (1)$$

The set of languages is G, and l_G denotes the number of languages. A corpus C in language g is denoted by C_g. A language model O based on C_g is denoted by $O(C_g)$. The features given values by the model $O(C_g)$ are the domain $dom(O(C_g))$ of the model. In a language model, a value v for the feature f is denoted by $v_{C_g}(f)$. For each potential language g of a corpus C in an unknown language, a resulting score $R(g, C)$ is calculated. A corpus in an unknown language is also referred to as a test document.

4.1 An Archetypal Language Identifier

The design of a supervised language identifier can generally be deconstructed into four key steps:

1. A representation of text is selected
2. A model for each language is derived from a training corpus of labelled documents
3. A function is defined that determines the similarity between a document and each language
4. The language of a document is predicted based on the highest-scoring model

4.2 On the Equivalence of Methods

The theoretical description of some of the methods leaves room for interpretation on how to implement them. Cormen et al. (1990) define an algorithm to be any well-defined computational procedure. Yanofsky (2011) introduces a three-tiered classification where programs implement algorithms and algorithms implement functions. The examples of functions given by Yanofsky (2011), *sort* and *find max* differ from our *identify language* as they are always solvable and produce the same results. In this survey, we have considered two methods to be the same if they always produce exactly the same results from exactly the same inputs. This would not be in line with the definition of an algorithm by Yanofsky (2011), as in his example there are two different algorithms *mergesort* and *quicksort* that implement the function *sort*, always producing identical results with the same input. What we in this survey call a method, is actually a function in the tiers presented by Yanofsky (2011).

5. Features

In this section, we present an extensive list of features used in LI, some of which are not self-evident. The equations written in the unified notation defined earlier show how the values v used in the language models are calculated from the tokens u. For each feature type, we generally introduce the first published article that used that feature type, as well as more recent articles where the feature type has been considered.

5.1 Bytes and Encodings

In LI, text is typically modeled as a stream of characters. However, there is a slight mismatch between this view and how text is actually stored: documents are digitized using a particular encoding, which is a mapping from characters (e.g. a character in an alphabet), onto the actual sequence of bytes that is stored and transmitted by computers. Encodings vary in how many bytes they use to represent each character. Some encodings use a fixed number of bytes for each character (e.g. ASCII), whereas others use a variable-length encoding (e.g. UTF-8). Some encodings are specific to a given language (e.g. GuoBiao 18030 or Big5 for Chinese), whereas others are specifically designed to represent as many languages as possible (e.g. the Unicode family of encodings). Languages can often be represented in a number of different encodings (e.g. UTF-8 and Shift-JIS for Japanese), and sometimes encodings are specifically designed to share certain codepoints (e.g. all single-byte UTF-8 codepoints are exactly the same as ASCII). Most troubling for LI, isomorphic encodings can be used to encode different languages, meaning that the determination of the encoding often doesn't help in honing in on the language. Infamous examples of this are the ISO-8859 and EUC encoding families. Encodings pose unique challenges for practical LI applications: a given language can often be encoded in different forms, and a given encoding can often map onto multiple languages.

Some LI research has included an explicit encoding detection step to resolve bytes to the characters they represent (Kikui, 1996), effectively transcoding the document into a standardized encoding before attempting to identify the language. However, transcoding is computationally expensive, and other research suggests that it may be possible to ignore encoding and build a single per-language model covering multiple encodings simultaneously (Kruengkrai et al., 2005; Baldwin and Lui, 2010b). Another solution is to treat each language-encoding pair as a separate category (Cowie et al., 1999; Suzuki et al., 2002; Singh and Gorla, 2007; Brown, 2012). The disadvantage of this is that it increases the computational cost by modeling a larger number of classes. Most of the research has avoided issues of encoding entirely by assuming that all documents use the same encoding (Mandl et al., 2006). This may be a reasonable assumption in some settings, such as when processing data from a single source (e.g. all data from Twitter and Wikipedia is UTF-8 encoded). In practice, a disadvantage of this approach may be that some encodings are only applicable to certain languages (e.g. S-JIS for Japanese and Big5 for Chinese), so knowing that a document is in a particular encoding can provide information that would be lost if the document is transcoded to a universal encoding such as UTF-8. Li and Momoi (2001) used a parallel state machine to detect which encoding scheme a file could potentially have been encoded with. The knowledge of the encoding, if detected, is then used to narrow down the possible languages.

Most features and methods do not make a distinction between bytes or characters, and because of this we will present feature and method descriptions in terms of characters, even if byte tokenization was actually used in the original research.

5.2 Characters

In this section, we review how individual character tokens have been used as features in LI.

Non-alphabetic or non-ideographic characters Ranaivo-Malançon and Ng (2005) used the formatting of numbers when distinguishing between Malay and Indonesian. King and Abney (2013) used the presence of non-alphabetic characters between the current word and the words before and after as features. Elfardy and Diab (2013) used emoticons (or emojis) in Arabic dialect identification with Naive Bayes ("NB"; see Section 6.5). Non-alphabetic characters have also been used by Basile et al. (2017), Bestgen (2017), Samih (2017), and Simaki et al. (2017).

Alphabets Henrich (1989) used knowledge of alphabets to exclude languages where a language-unique character in a test document did not appear. Giguet (1995) used alphabets collected from dictionaries to check if a word might belong to a language. Hanif et al. (2007) used the Unicode database to get the possible languages of individual Unicode characters. Lately, the knowledge of relevant alphabets has been used for LI also by Hasimu and Silamu (2017) and Samih (2017).

Capitalization Capitalization is mostly preserved when calculating character n-gram frequencies, but in contexts where it is possible to identify the orthography of a given document and where capitalization exists in the orthography, lowercasing can be used to reduce sparseness. In recent LI work, capitalization was used as a special feature by Basile et al. (2017), Bestgen (2017), and Simaki et al. (2017).

The number of characters in words and word combinations Langer (2001) was the first to use the length of words in LI. Nobesawa and Tahara (2005) used the length of full person names comprising several words. Lately, the number of characters in words has been used for LI by Dongen (2017), van der Lee and Bosch (2017), Samih (2017), and Simaki et al. (2017). Dongen (2017) also used the length of the two preceding words.

The frequency or probability of each character Kerwin (2006) used character frequencies as feature vectors. In a feature vector, each feature f has its own integer value. The raw frequency – also called term frequency (TF) – is calculated for each language g as:

$$v_{C_g}(f) = c(C_g, f) \tag{2}$$

Rau (1974) was the first to use the probability of characters. He calculated the probabilities as relative frequencies, by dividing the frequency of a feature found in the corpus by the total count of features of the same type in the corpus. When the relative frequency of a feature f is used as a value, it is calculated for each language g as:

$$v_{C_g}(f) = \frac{c(C_g, f)}{l_{C_g^F}} \tag{3}$$

Tran and Sharma (2005) calculated the relative frequencies of one character prefixes, and Windisch and Csink (2005) did the same for one character suffixes.

Ng and Selamat (2009) calculated character frequency document frequency ("LFDF") values. Takçı and Güngör (2012) compared their own Inverse Class Frequency ("ICF") method with the Arithmetic Average Centroid ("AAC") and the Class Feature Centroid ("CFC") feature vector updating methods. In ICF a character appearing frequently only in some language gets more positive weight for that language. The values differ from Inverse Document Frequency ("IDF", Equation 8), as they are calculated using also the frequencies of characters in other languages. Their ICF-based vectors generally performed better than those based on AAC or CFC. Takçı and Ekinci (2012) explored using the relative frequencies of characters with similar discriminating weights. Takçı and Güngör (2012) also used Mutual Information ("MI") and chi-square weighting schemes with characters.

Baldwin and Lui (2010b) compared the identification results of single characters with the use of character bigrams and trigrams when classifying over 67 languages. Both bigrams and trigrams generally performed better than unigrams. Jauhiainen (2010) also found that the identification results from identifiers using just characters are generally worse than those using character sequences.

5.3 Character Combinations

In this section we consider the different combinations of characters used in the literature. Character n-grams mostly consist of all possible characters in a given encoding, but can also consist of only alphabetic or ideographic characters.

Co-occurrences Windisch and Csink (2005) calculated the co-occurrence ratios of any two characters, as well as the ratio of consonant clusters of different sizes to the total number of consonants. Sterneberg (2012) used the combination of every bigram and their counts in words. van der Lee and Bosch (2017) used the proportions of question and exclamation marks to the total number of the end of sentence punctuation as features with several machine learning algorithms.

Franco-Salvador et al. (2017b) used FastText to generate character n-gram embeddings (Joulin et al., 2017). Neural network generated embeddings are explained in Section 5.6.

Vowel-consonant relationship Rau (1974) used the relative frequencies of vowels following vowels, consonants following vowels, vowels following consonants and consonants following consonants. Dongen (2017) used vowel-consonant ratios as one of the features with Support Vector Machines ("SVMs", Section 6.8), Decision Trees ("DTs", Section 6.2), and Conditional Random Fields ("CRFs", Section 10.7).

Character repetition Elfardy and Diab (2013) used the existence of word lengthening effects and repeated punctuation as features. Banerjee et al. (2014) used the presence of characters repeating more than twice in a row as a feature with simple scoring (Equation 17). Barman et al. (2014a) used more complicated repetitions identified by regular expressions. Sikdar and Gambäck (2016) used letter and character bigram repetition with a CRF. Martinc et al. (2017) used the count of character sequences with three or more identical characters, using several machine learning algorithms.

Article	1	2	3	4	5	6	7
Bošnjak et al. (2013)				***			
Zampieri et al. (2013)		*	**	***	*		
Das and Gambäck (2014)	*	*	**	**	*	*	*
Indhuja et al. (2014)	***						
Ljubešić and Kranjcić (2014)			*			*	
Minocha and Tyers (2014)	*	*	***	*	*		
Pethö and Mózes (2014)		*	*	**	***		
Sadat et al. (2014b,a)	*	***	*				
Tan et al. (2014)		*	*	*	***	**	
Zaidan and Callison-Burch (2014)	*		**		***		
Zampieri and Gebre (2014)			***				
Franco-Salvador et al. (2015a)				***			
Jauhiainen et al. (2015b)	*	*	*	*	*	***	**...
King et al. (2015)					***		
Panich (2015)	*	*	*	*	*	*	
Zampieri et al. (2015a)					***		
Abainia et al. (2016)	*	*	*				
Castro et al. (2016, 2017)		*	*	*	*	***	**
Giwa (2016)	*	*	**	***	*	*	*
Hanani et al. (2016)			***				
Duvenhage et al. (2017)					***		
Jourlin (2017)	***						

Table 2: List of articles (2013–2017) where relative frequencies of character n-grams have been used as features. The columns indicate the length of the n-grams used. "***" indicates the empirically best n-gram length in that paper, and "**" the second-best n-gram length. "*" indicates that there was no clear winner in terms of n-gram order reported on in the paper.

n-grams of characters of the same size Character n-grams are continuous sequences of characters of length n. They can be either consecutive or overlapping. Consecutive character bigrams created from the four character sequence *door* are *do* and *or*, whereas the overlapping bigrams are *do*, *oo*, and *or*. Overlapping n-grams are most often used in the literature. Overlapping produces a greater number and variety of n-grams from the same amount of text.

Rau (1974) was the first to use combinations of any two characters. He calculated the relative frequency of each bigram. Table 2 lists more recent articles where relative frequencies of n-grams of characters have been used. Rau (1974) also used the relative frequencies of two character combinations which had one unknown character between them, also known as gapped bigrams. Seifart and Mundry (2015) used a modified relative frequency of character unigrams and bigrams.

Character trigram frequencies relative to the word count were used by Vega and Bressan (2001a), who calculated the values $v_C(f)$ as in Equation 4. Let T be the word-tokenized segmentation of the corpus C of character tokens, then:

$$v_C(f) = \frac{c(C, f)}{l_T} \tag{4}$$

Article	2	3	4	5	6	7	8
Ramisch (2008)	**	**	**	***			
You et al. (2008)		**	***	**	*	*	
Stupar et al. (2011); Tiedemann and Ljubešić (2012)		***					
Goldszmidt et al. (2013)		***					
Bar and Dershowitz (2014)		***					
Brown (2014)		***					
Gamallo et al. (2014)		***					
Hurtado et al. (2014)			***				
Indhuja et al. (2014)	*	*					
Leidig (2014)				***			
Mendizabal et al. (2014)	*	*	**	***	**	**	**
Pethö and Mózes (2014)	*	*	**	***			
Sadat et al. (2014b,a)	***	**					
Ullman (2014)			***				
Cianflone and Kosseim (2016)	*	*	*	*	*	**	***
Martadinata et al. (2016)		*	**	***			
Samih and Maier (2016); Samih (2017)				***			

Table 3: List of recent articles where Markovian character n-grams have been used as features. The columns indicate the length of the n-grams used. "***" indicates the best and "**" the second-best n-gram length as reported in the article in question. "*" indicates that there was no clear winner in terms of n-gram order reported on in the paper.

where $c(C, f)$ is the count of character trigrams f in C, and l_T is the total word count in the corpus. Later n-gram frequencies relative to the word count were used by Hamzah (2010) for character bigrams and trigrams.

House and Neuburg (1977) divided characters into five phonetic groups and used a Markovian method to calculate the probability of each bigram consisting of these phonetic groups. In Markovian methods, the probability of a given character u_i is calculated relative to a fixed-size character context $u_{i-n+1}, ..., u_{i-1}$ in corpus C, as follows:

$$P(u_i|u_{i-n+1,...,i-1}) = \frac{c(C, u_{i-n+1,...,i})}{c(C, u_{i-n+1,...,i-1})} \tag{5}$$

where $u_{i-n+1,...,i-1}$ is an n-gram prefix of $u_{i-n+1,...,i}$ of length $n-1$. In this case, the probability $P(u_i|u_{i-n+1,...,i-1})$ is the value $v_C(f)$, where $f = u_{i-n+1,...,i}$, in the model $O(C)$. Ludovik and Zacharski (1999) used 4-grams with recognition weights which were derived from Markovian probabilities. Table 3 lists some of the more recent articles where Markovian character n-grams have been used.

Vitale (1991) was the first author to propose a full-fledged probabilistic language identifier. He defines the probability of a trigram f being written in the language g to be:

$$P(g|f) = \frac{P(f|g)P(g)}{\sum_{h \in G} P(f|h)P(h)} \tag{6}$$

He considers the prior probabilities of each language $P(g)$ to be equal, which leads to:

$$P(g|f) = \frac{P(f|g)}{\sum_{h \in G} P(f|h)} \tag{7}$$

Article	1	2	3	4	5	6+	Method
Adouane (2016)	*	*	*	*	*	*	SVM (cf. §6.8)
Al-Badrashiny and Diab (2016)	*	*	*	*	*		CRF (cf. §10.7)
Alshutayri et al. (2016)	*	*	*				SVM, DT (cf. §6.2)
Barbaresi (2016)		*	*	*	*	*	NB (cf. §6.5), XGBoost (cf. §6.10), RF (cf. §6.2)
Ciobanu and Dinu (2016)	*	*	*	*			LR (cf. §6.7)
Giwa (2016)		*	*	*	*		SVM
Goutte and Léger (2016)						*	SVM, NB
Ionescu and Popescu (2016)		*	*	*	*	*	KRR (cf. §6.6), KDA (cf. §6.6)
Lamabam and Chakma (2016)			*				CRF
Criscuolo and Aluísio (2017)					*		NB
Malmasi and Dras (2017)	*	*	*	*	*	*	SVM
Malmasi (2017)				*			SVM
Mathur et al. (2017)	*	*	*	*	*	*	NB, LR
Oliveira and Neto (2017)			*	*	*	*	SVM
Plaza Cagigós (2017)			*				NB, SVM, DT, NN (cf. §6.9)
Rangel et al. (2017a)				*			SVM, NB, NN, ...
Schaetti (2017)		*					NN

Table 4: Recent papers (2016–) where the frequency of character n-grams has been used to generate feature vectors. The columns indicate the length of the n-grams used, and the machine learning method(s) used. The relevant section numbers for the methods are mentioned in parentheses.

Vitale (1991) used the probabilities $P(g|f)$ as the values $v_{C_g}(f)$ in the language models.

MacNamara et al. (1998) used a list of the most frequent bigrams and trigrams with logarithmic weighting. Prager (1999) was the first to use direct frequencies of character n-grams as feature vectors. Vinosh Babu and Baskaran (2005) used Principal Component Analysis ("PCA") to select only the most discriminating bigrams in the feature vectors representing languages. Murthy and Kumar (2006) used the most frequent and discriminating byte unigrams, bigrams, and trigrams among their feature functions. They define the most discriminating features as those which have the most differing relative frequencies between the models of the different languages. Gottron and Lipka (2010) tested n-grams from two to five using frequencies as feature vectors, frequency ordered lists, relative frequencies, and Markovian probabilities. Table 4 lists the more recent articles where the frequency of character n-grams have been used as features. In the method column, "RF" refers to Random Forest (cf. Section 6.2), "LR" to Logistic Regression (Section 6.7), "KRR" to Kernel Ridge Regression (Section 6.6), "KDA" to Kernel Discriminant Analysis (Section 6.6), and "NN" to Neural Networks (Section 6.9).

Giguet (1995) used the last two and three characters of open class words. Suzuki et al. (2002) used an unordered list of distinct trigrams with the simple scoring method (Section 6.3). Hayati (2004) used Fisher's discriminant function to choose the 1000 most discriminating trigrams. Bilcu and Astola (2006) used unique 4-grams of characters with positive Decision Rules (Section 6.1). Ozbek et al. (2006) used the frequencies of bi- and trigrams in words unique to a language. Milne et al. (2012) used lists of the most frequent trigrams.

Li and Momoi (2001) divided possible character bigrams into those that are commonly used in a language and to those that are not. They used the ratio of the commonly used bigrams to all observed bigrams to give a confidence score for each language. Xafopoulos et al. (2004) used the difference between the ISO Latin-1 code values of two consecutive characters as well as two characters separated by another character, also known as gapped character bigrams.

Artemenko and Shramko (2005) used the IDF and the transition probability of trigrams. They calculated the IDF values $v_{C_g}(f)$ of trigrams f for each language g, as in Equation 8, where $c(C_g, f)$ is the number of trigrams f in the corpus of the language g and $df(C_G, f)$ is the number of languages in which the trigram f is found, where C_G is the language-segmented training corpus with each language as a single segment.

$$v_{C_g}(f) = \frac{c(C_g, f)}{df(C_G, f)} \tag{8}$$

df is defined as:

$$df(C_G, f) = \sum_{g \in G} \begin{cases} 1 & \text{, if } c(C_g, f) > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{9}$$

Malmasi et al. (2015) used n-grams from one to four, which were weighted with "TF-IDF" (Term Frequency–Inverse Document Frequency). TF-IDF was calculated as:

$$v_{C_g}(f) = c(C_g, f) \log \frac{l_G}{df(C_G, f)} \tag{10}$$

TF-IDF weighting or close variants have been widely used for LI. Thomas and Verma (2007) used "CF-IOF" (Class Frequency-Inverse Overall Frequency) weighted 3- and 4-grams.

Jhamtani et al. (2014) used the logarithm of the ratio of the counts of character bigrams and trigrams in the English and Hindi dictionaries. Zamora et al. (2014) used a feature weighting scheme based on mutual information ("MI"). They also tried weighting schemes based on the "GSS" (Galavotti, Sebastiani, and Simi) and "NGL" (Ng, Goh, and Low) coefficients, but using the MI-based weighting scheme proved the best in their evaluations when they used the sum of values method (Equation 18). Martinc et al. (2017) used punctuation trigrams, where the first character has to be a punctuation mark (but not the other two characters). Saharia (2017) used consonant bi- and trigrams which were generated from words after the vowels had been removed.

Character n-grams of differing sizes The language models mentioned earlier consisted only of n-grams of the same size n. If n-grams from one to four were used, then there were four separate language models. Cavnar and Trenkle (1994) created ordered lists of the most frequent n-grams for each language. Singh and Goyal (2014) used similar n-gram lists with symmetric cross-entropy. Russell and Lapalme (2003) used a Markovian method to calculate the probability of byte trigrams interpolated with byte unigrams. Vatanen et al. (2010) created a language identifier based on character n-grams of different sizes over 281 languages, and obtained an identification accuracy of 62.8% for extremely short samples (5–9 characters). Their language identifier was used or evaluated by Rodrigues (2012),

Maier and Gómez-Rodríguez (2014), and Jauhiainen et al. (2017a). Rodrigues (2012) managed to improve the identification results by feeding the raw language distance calculations into an SVM.

Table 5 lists recent articles where character n-grams of differing sizes have been used. "LR" in the methods column refer to Logistic Regression (Section 6.6), "LSTM RNN" to Long Short-Term Memory Recurrent Neural Networks (Section 6.9), and "DAN" to Deep Averaging Networks (Section 6.9). Kikui (1996) used up to the four last characters of words and calculated their relative frequencies. Ahmed et al. (2004) used frequencies of 2–7-grams, normalized relative to the total number of n-grams in all the language models as well as the current language model. Jauhiainen (2010) compared the use of different sizes of n-grams in differing combinations, and found that combining n-grams of differing sizes resulted in better identification scores. Lui and Baldwin (2011, 2012, 2014) used mixed length domain-independent language models of byte n-grams from one to three or four.

Mixed length language models were also generated by Brown (2012) and later by Brown (2013, 2014), who used the most frequent and discriminating n-grams longer than two bytes, up to a maximum of 12 bytes, based on their weighted relative frequencies. K of the most frequent n-grams were extracted from training corpora for each language, and their relative frequencies were calculated. In the tests reported in (Brown, 2013), K varied from 200 to 3,500 n-grams. Later Sanchez-Perez et al. (2017) also evaluated different combinations of character n-grams as well as their combinations with words.

Stensby et al. (2010) used mixed-order n-gram frequencies relative to the total number of n-grams in the language model. Sterneberg (2012) used frequencies of n-grams from one to five and gapped 3- and 4-grams as features with an SVM. As an example, some gapped 4-grams from the word *Sterneberg* would be *Senb*, *tree*, *enbr*, and *reeg*. King et al. (2014b) used character n-grams as a backoff from Markovian word n-grams. Shrestha (2014) used the frequencies of word initial n-grams ranging from 3 to the length of the word minus 1. Ács et al. (2015) used the most relevant n-grams selected using the absolute value of the Pearson correlation. Mandal et al. (2015) used only the first 10 characters from a longer word to generate the n-grams, while the rest were ignored. Qiao and Lévy (2015) used only those n-grams which had the highest TF-IDF scores. Bestgen (2017) used character n-grams weighted by means of the "BM25" (Best Match 25) weighting scheme. Hanani et al. (2017) used byte n-grams up to length 25.

Consonant or vowel sequences Sterneberg (2012) used consonant sequences generated from words. Anand (2014) used the presence of vowel sequences as a feature with a NB classifier (see Section 6.5) when distinguishing between English and transliterated Indian languages.

n-gram dictionary Chanda et al. (2016a) used a basic dictionary (Section 5.5) composed of the 400 most common character 4-grams.

Unique character combinations Henrich (1989) and Vitale (1991) used character combinations (of different sizes) that either existed in only one language or did not exist in one or more languages.

Article	1	2	3	4	5	6	7	8+	Method
Adouane et al. (2016b,a)	*	*	*						SVM (cf. §6.8)
Adouane et al. (2016c,d)					*	*			SVM
Balažević et al. (2016)	*	*	*	*					Product (cf. §6.5), SVM, LR (cf. §6.7), Sum (cf. §6.4)
Çöltekin and Rama (2016)	*	*	*	*	*	*			SVM
Eldesouki et al. (2016)		*	*	*	*				SVM, LR, NN (cf. §6.9), NB (cf. §6.5)
He et al. (2016)	*	*	*	*					LR (cf. §6.6)
Jauhiainen et al. (2016)	*	*	*	*	*	*	*	*	HeLI (cf. §6.5)
Malmasi and Zampieri (2016, 2017a)	*	*	*	*	*	*			SVM
Piergallini et al. (2016b)	*	*	*	*					LR
Piergallini et al. (2016a)	*	*	*						LR
Radford and Gallé (2016)		*	*	*	*				LR
Samih et al. (2016)		*	*						LSTM RNN (cf. §6.9)
Xu et al. (2016)	*	*	*						SVM
Alrifai et al. (2017)		*	*	*	*	*	*		SVM
Barbaresi (2017)		*	*	*	*	*	*		NB
Bestgen (2017)	*	*	*	*	*	*	*	*	SVM
Clematide and Makarov (2017)	*	*	*	*	*	*			NB, CRF (cf. §10.7), SVM
Espichán-Linares and Oncevay-Marcos (2017)		*	*						SVM, NB, ...
Franco-Salvador et al. (2017b)			*	*	*	*			DAN (cf. §6.9)
Gamallo et al. (2017)					*	*	*		Perplexity (cf. §6.6)
Gómez-Adorno et al. (2017)			*	*	*				NB, SVM
Hanani et al. (2017)	*	*	*						SVM, NB, LR, DT (cf. §6.2)
Jauhiainen et al. (2017b)	*	*	*	*	*	*	*	*	HeLI
Jauhiainen et al. (2017a)	*	*	*	*	*	*			HeLI
Malmasi and Dras (2017)	*	*	*	*	*	*			SVM
Malmasi and Zampieri (2017b)	*	*	*	*	*	*	*	*	SVM
Mathur et al. (2017)		*	*	*	*				RNN
Miura et al. (2017)	*	*	*	*	*	*			SVM
Sanchez-Perez et al. (2017)			*	*	*	*	*	*	SVM
Tellez et al. (2017)	*		*		*		*	*	SVM
Espichán-Linares and Oncevay-Marcos (2018)		*	*	*					SVM, NB, ...

Table 5: List of articles (2016-) where character n-grams of differing sizes have been used as features. The numbered columns indicate the length of the n-grams used. The method column indicates the method used with the n-grams. The relevant section numbers are mentioned in parentheses.

5.4 Morphemes, Syllables and Chunks

Morphemes Giguet (1998) used the suffixes of lexical words derived from untagged corpora. El-Shishiny et al. (2004) used prefixes and suffixes determined using linguistic knowledge of the Arabic language. Marcadet et al. (2005) used suffixes and prefixes in rule-based LI. Ozbek et al. (2006) used morphemes and morpheme trigrams (morphotactics)

Reference	1	2	3	4	Method
He et al. (2016)	*	*	*		LR (cf. §6.6)
Piergallini et al. (2016b)	*	*	*	*	LR
Samih and Maier (2016); Samih (2017)	*	*	*		CRF (cf. §10.7)
Schulz and Keller (2016)	*	*	*		CRF
Shrestha (2016)	*	*	*	*	CRF
Sikdar and Gambäck (2016)	*	*	*	*	CRF
Xia (2016)	*	*	*		CRF
Clematide and Makarov (2017)	*	*	*		CRF
Gómez-Adorno et al. (2017)			*		NB (cf. §6.5), SVM (cf. §6.8)
Martinc et al. (2017)				*	SVM, LR, RF (cf. §6.2), ...

Table 6: References (2016-) where prefixes and suffixes collected from a training corpus have been used for LI. The columns indicate the length of the prefixes and suffixes. The method column indicates the method used. The relevant section numbers are mentioned in parentheses.

constructed by Creutz's algorithm (Creutz, 2003). Hammarström (2007) used prefixes and suffixes constructed by his own algorithm, which was later also used by Ceylan and Kim (2009). Romsdorfer and Pfister (2007) used morpheme lexicons in LI. Ceylan and Kim (2009) compared the use of morphological features with the use of variable sized character n-grams. When choosing between ten European languages, the morphological features obtained only 26.0% accuracy while the n-grams reached 82.7%. Yeong and Tan (2010) lemmatized Malay words in order to get the base forms. Lu and Mohamed (2011) used a morphological analyzer of Arabic. Zampieri et al. (2013) used morphological information from a part-of-speech (POS) tagger. Anand (2014) and Banerjee et al. (2014) used manually selected suffixes as features. Bekavac et al. (2014) created morphological grammars to distinguish between Croatian and Serbian. Darwish et al. (2014) used morphemes created by Morfessor, but they also used manually created morphological rules. Gamallo et al. (2014) used a suffix module containing the most frequent suffixes. Dutta et al. (2015) and Mandal et al. (2015) used word suffixes as features with CRFs. Barbaresi (2016) used an unsupervised method to learn morphological features from training data. The method collects candidate affixes from a dictionary built using the training data. If the remaining part of a word is found from the dictionary after removing a candidate affix, the candidate affix is considered to be a morpheme. Barbaresi (2016) used 5% of the most frequent affixes in language identification. Gómez-Adorno et al. (2017) used character n-grams classified into different types, which included prefixes and suffixes. Table 6 lists some of the more recent articles where prefixes and suffixes collected from a training corpus has been used for LI.

Syllables and syllable n-grams Chen et al. (2006) used trigrams composed of syllables. Yeong and Tan (2010) used Markovian syllable bigrams for LI between Malay and English. Later Yeong and Tan (2011) also experimented with syllable uni- and trigrams. Murthy and Kumar (2006) used the most frequent as well as the most discriminating Indian script syllables, called aksharas. They used single aksharas, akshara bigrams, and akshara trigrams. Syllables would seem to be especially apt in situations where distinction needs to be made between two closely-related languages.

Chunks, chunk n-grams and n-grams of n-grams You et al. (2008) used the trigrams of non-syllable chunks that were based on MI. Yeong and Tan (2010) experimented also with Markovian bigrams using both character and grapheme bigrams, but the syllable bigrams proved to work better. Graphemes in this case are the minimal units of the writing system, where a single character may be composed of several graphemes (e.g. in the case of the Hangul or Thai writing systems). Later, Yeong and Tan (2011) also used grapheme uni- and trigrams. Yeong and Tan (2011) achieved their best results combining word unigrams and syllable bigrams with a grapheme back-off. Elfardy et al. (2014) used the MADAMIRA toolkit for D3 decliticization and then used D3-token 5-grams. D3 decliticization is a way to preprocess Arabic words presented by Habash and Sadat (2006).

Graphones are sequences of characters linked to sequences of corresponding phonemes. They are automatically deduced from a bilingual corpus which consists of words and their correct pronunciations using Joint Sequence Models ("JSM"). Giwa and Davel (2014) used language tags instead of phonemes when generating the graphones and then used Markovian graphone n-grams from 1 to 8 in LI.

5.5 Words

Position of words Kumar et al. (2015) used the position of the current word in word-level LI. The position of words in sentences has also been used as a feature in code-switching detection by Dongen (2017). It had predictive power greater than the language label or length of the previous word.

The characteristics of words Mustonen (1965) used the characteristics of words as parts of discriminating functions. Barman et al. (2014b) used the string edit distance and n-gram overlap between the word to be identified and words in dictionaries. Similarly Jhamtani et al. (2014) used a modified edit distance, which considers the common spelling substitutions when Hindi is written using latin characters. Das and Gambäck (2013) used the Minimum Edit Distance ("MED").

Basic dictionary Basic dictionaries are unordered lists of words belonging to a language. Basic dictionaries do not include information about word frequency, and are independent of the dictionaries of other languages. Vitale (1991) used a dictionary for LI as a part of his speech synthesizer. Each word in a dictionary had only one possible "language", or pronunciation category. More recently, a basic dictionary has been used for LI by Adouane and Dobnik (2017), Dongen (2017), and Duvenhage et al. (2017).

Dictionary of unique words Unique word dictionaries include only those words of the language, that do not belong to the other languages targeted by the language identifier. Kulikowski (1991) used unique short words (from one to three characters) to differentiate between languages. Recently, a dictionary of unique words was used for LI by Adouane (2016), Guellil and Azouaou (2016), and Martinc et al. (2017).

Specific classes of words Giguet (1995) used exhaustive lists of function words collected from dictionaries. Wechsler et al. (1997) used stop words – that is non-content or closed-class words – as a training corpus. Similarly, Lins and Gonçalves (2004) used words from closed word classes, and Stupar et al. (2011) used lists of function words.

Al-Badrashiny et al. (2015) used a lexicon of Arabic words and phrases that convey modality. Common to these features is that they are determined based on linguistic knowledge.

Discriminating words Rehůřek and Kolkus (2009) used the most relevant words for each language. Babu and Kumar (2010) used unique or nearly unique words. Franco-Salvador et al. (2015a) used Information Gain Word-Patterns ("IG-WP") to select the words with the highest information gain.

Most common words Souter et al. (1994) made an (unordered) list of the most common words for each language, as, more recently, did Cazamias et al. (2015), Panich (2015), and Abainia et al. (2016). Pavan et al. (2010) encoded the most common words to root forms with the Soundex algorithm.

Word frequency Mather (1998) collected the frequencies of words into feature vectors. Prager (1999) compared the use of character n-grams from 2 to 5 with the use of words. Using words resulted in better identification results than using character bigrams (test document sizes of 20, 50, 100 or 200 characters), but always worse than character 3-, 4- or 5-grams. However, the combined use of words and character 4-grams gave the best results of all tested combinations, obtaining 95.6% accuracy for 50 character sequences when choosing between 13 languages. Ács et al. (2015) used TF-IDF scores of words to distinguish between language groups. Recently, the frequency of words has also been used for LI by Clematide and Makarov (2017), Gómez-Adorno et al. (2017), Plaza Cagigós (2017), and Saharia (2017).

The relative frequency of words Poutsma (2002) and Zhdanova (2002) were the first to use relative frequencies of words in LI. As did Prager (1999) for word frequencies, also Jauhiainen (2010) found that combining the use of character n-grams with the use of words provided the best results. His language identifier obtained 99.8% average recall for 50 character sequences for the 10 evaluated languages (choosing between the 13 languages known by the language identifier) when using character n-grams from 1 to 6 combined with words. Tiedemann and Ljubešić (2012) calculated the relative frequency of words over all the languages. Artemenko and Shramko (2005) calculated the IDF of words, following the approach outlined in Equation 8. Xu et al. (2016) calculated the Pointwise Mutual Information ("PMI") for words and used it to group words to Chinese dialects or dialect groups. Recently, the relative frequency of words has also been used for LI by Jauhiainen et al. (2017b,a) and Jourlin (2017)

Short words Grefenstette (1995) used the relative frequency of words with less than six characters. Recently, Panich (2015) also used short words, as did Simaki et al. (2017).

Search engine queries Alex (2005) used the relative frequency calculated from Google searches. Google was later also used by You et al. (2008) and Yang and Liang (2010).

Word probability maps Scherrer and Rambow (2010) created probability maps for words for German dialect identification between six dialects. In a word probability map, each predetermined geographic point has a probability for each word form. Probabilities were derived using a linguistic atlas and automatically-induced dialect lexicons.

Morphological analyzers and spellchecking Pienaar and Snyman (2010) used commercial spelling checkers, which utilized lexicons and morphological analyzers. The language identifier of Pienaar and Snyman (2010) obtained 97.9% accuracy when classifying one-line texts between 11 official South African languages. Elfardy and Diab (2012) used the AL-MORGEANA analyzer to check if the word had an analysis in modern standard Arabic. They also used sound change rules to use possible phonological variants with the analyzer. Joshi et al. (2013) used spellchecking and morphological analyzers to detect English words from Hindi–English mixed search queries. Akosu and Selamat (2014) used spelling checkers to distinguish between 15 languages, extending the work of Pienaar and Snyman (2010) with dynamic model selection in order to gain better performance. Shrestha (2014) used a similarity count to find if mystery words were misspelled versions of words in a dictionary.

Word clusters Pham and Tran (2003) used an "LBG-VQ" (Linde, Buzo & Gray algorithm for Vector Quantization) approach to design a codebook for each language (Linde et al., 1980). The codebook contained a predetermined number of codevectors. Each codeword represented the word it was generated from as well as zero or more words close to it in the vector space.

5.6 Word Combinations

Sentence length Elfardy and Diab (2013) used the number of words in a sentence with NB. van der Lee and Bosch (2017) and Simaki et al. (2017) used the sentence length calculated in both words and characters with several machine learning algorithms.

Statistics of words van der Lee and Bosch (2017) used the ratio to the total number of words of: once-occurring words, twice-occurring words, short words, long words, function words, adjectives and adverbs, personal pronouns, and question words. They also used the word-length distribution for words of 1–20 characters.

Word n-grams Marcadet et al. (2005) used at least the preceding and proceeding words with manual rules in word-level LI for text-to-speech synthesis. Rosner and Farrugia (2007) used Markovian word n-grams with a Hidden Markov Model ("HMM") tagger (Section 6.10). Table 7 lists more recent articles where word n-grams or similar constructs have been used. "PPM" in the methods column refers to Prediction by Partial Matching (Section 5.7), and "kNN" to k Nearest Neighbor classification (Section 6.11).

Singh (2006) used word trigrams simultaneously with character 4-grams. He concluded that word-based models can be used to augment the results from character n-grams when they are not providing reliable identification results. Table 8 lists articles where both character and word n-grams have been used together. "CBOW" in the methods column refer to Continuous Bag of Words neural network (Section 6.9), and "MIRA" to Margin Infused Relaxed Algorithm (Section 6.8). Sanchez-Perez et al. (2017) evaluated different combinations of word and character n-grams with SVMs. The best combination for language variety identification was using all the features simultaneously. Tellez et al. (2017) used normal and gapped word n-grams and character n-grams simultaneously.

Co-occurrences of words Wan (2016) uses word embeddings consisting of Positive Pointwise Mutual Information ("PPMI") counts to represent each word type. Then they

use Truncated Singular Value Decomposition ("TSVD") to reduce the dimension of the word vectors to 100. Elgabou and Kazakov (2017) used k-means clustering when building dialectal Arabic corpora. Kheng et al. (2017) used features provided by Latent Semantic Analysis ("LSA") with SVMs and NB.

Mikolov et al. (2013) present two models, the CBOW model and the continuous skip-gram model. The CBOW model can be used to generate a word given it's context and the skip-gram model can generate the context given a word. The projection matrix, which is the weight matrix between the input layer and the hidden layer, can be divided into vectors, one vector for each word in the vocabulary. These word-vectors are also referred to as word embeddings. The embeddings can be used as features in other tasks after the neural network has been trained. Lin et al. (2014), Chang and Lin (2014), Franco-Salvador et al. (2015a,b), Jain (2015), Franco-Salvador et al. (2017a,b), and Rangel et al. (2017a) used word embeddings generated by the word2vec skip-gram model (Mikolov et al., 2013) as features in LI. Poulston et al. (2017) used word2vec word embeddings and k-means clustering. Akhtyamova et al. (2017), Kodiyan et al. (2017), and Samih (2017) also used word embeddings created with word2vec.

Çöltekin and Rama (2016) trained both character and word embeddings using FastText text classification method (Joulin et al., 2017) on the Discriminating between Similar Languages ("DSL") 2016 shared task, where it reached low accuracy when compared with the other methods. Xia (2016) used FastText to train word vectors including subword information. Then he used these word vectors together with some additional word features to train a CRF-model which was used for codeswitching detection.

Article	1	2	3	4	5	6	7	Method
Bhattu and Ravi (2015)	***	***	***	***	***			LR (cf. §6.6)
Bobicev (2015)		***						PPM-C (cf. §5.7)
Ghosh et al. (2015)	***	***	***	***	***	***	***	CRF (cf. §10.7)
Huang (2015)	***	*	*					Product (cf. §6.5)
Qiao and Lévy (2015)	*	*						SVM (cf. §6.8)
Raghavi et al. (2015)	***	***	***	***	***	***	***	SVM
Shah et al. (2015)		***	***	***	***			SVM
Adouane et al. (2016a,b)	***	**	**	*	*			SVM
Adouane (2016)	***	***	**	*				NB (cf. §6.5), SVM, LR, kNN (cf. §6.11), DT (cf. §6.2)
Barbaresi (2016)		***						RF (cf. §6.2)
Eldesouki et al. (2016)	***	*	*					SVM, LR, NN (cf. §6.9), NB
Franco-Penya and Sanchez (2016)	*	***						NB, SVM
Hanani et al. (2016)	***	***	***					LSTM RNN (cf. §6.9)
Samih and Maier (2016); Samih (2017)	***	***	***	***	***	***	***	CRF
Samih et al. (2016)	***	***	***	***	***			LSTM RNN - CRF
Schulz and Keller (2016)	***	***	***	***	***			CRF
Sikdar and Gambäck (2016)	***	***	***	***	***			CRF
Xia (2016)	***	***	***					CRF
Zampieri et al. (2016)	***	**						SVM
Alrifai et al. (2017)	*	*	*					SVM
Criscuolo and Aluísio (2017)	***	***						NN (cf. §6.9)
Gamallo et al. (2017)	***	***	***					Perplexity (cf. §6.6)
Hanani et al. (2017)	***	***	*					SVM
Kheng et al. (2017)	***	***	***					NB, SVM, RF
van der Lee and Bosch (2017)								AdaBoost (cf. §6.11), DT, SVM, NB, ...
Mathur et al. (2017)	**	***	*	*	*	*	*	NB, LR
Mendoza and Mendelsohn (2017)		***						Product
Miura et al. (2017)	***	***						SVM
Poulston et al. (2017)	***	***						LR
Rangel et al. (2017a)		***						SVM, NB, NN, ...
Rijhwani et al. (2017)	***	***	***					HMM (cf. §6.10)
Sanchez-Perez et al. (2017)	***	***						SVM
Tellez et al. (2017)	***	***	***					SVM

Table 7: References (2015–) where word n-grams have been used as features. The numbered columns indicate the length of the n-grams used. "***" indicates the best and "**" the second best n-gram length, as evaluated in the article in question. "*" indicates that there was no clear order of effectiveness, or that the order was not presented in the article. The method column indicates the method used. The relevant section numbers are mentioned in parentheses.

Barman et al. (2014b) extracted features from the hidden layer of a Recurrent Neural Network ("RNN") that had been trained to predict the next character in a string. They used the features with a SVM classifier.

Article	1	2	3	4	5	6	7	char	Method
Singh (2006, 2010)			✓					1-4	similarity measures (cf. §6.6)
Das and Gambäck (2013, 2014)	✓	✓	✓	✓	✓	✓	✓	1-7	SVM (cf. §6.8)
Nguyen and Dogruöz (2013)	✓	✓	✓					1-5	LR (cf. §6.6), CRF (cf. §10.7)
Chittaranjan et al. (2014)	✓	✓	✓					1-5	CRF
Darwish et al. (2014)	✓	✓	✓					1-5	RF (cf. §6.2)
Goutte et al. (2014)	✓	✓						2-6	SVM, NB-like (cf. §6.5)
Gupta et al. (2014)	✓	✓	✓	✓	✓			1-3	RF, SVM, DT (cf. §6.2), ...
King et al. (2014a)	✓	✓						1-5	NB, LR, SVM
Ullman (2014)	✓	✓						1-4	NB-like
Ács et al. (2015)	✓	✓						1-4	LR, SVM
Chang and Lin (2014)	✓	✓	✓					2-3	RNN (cf. §6.9)
Goutte and Léger (2015)	✓	✓						2-6	SVM, NB-like
Jain (2015)	✓	✓	✓	✓	✓	✓	✓	2-4	CRF
Malmasi and Dras (2015a)	✓	✓						1-3	SVM
Malmasi and Dras (2015b)	✓	✓						1-6	SVM
Malmasi et al. (2015)	✓	✓						1-4	SVM
Castro et al. (2016)	✓	✓						2-7	Product (cf. §6.5)
Zirikly et al. (2016)	✓	✓	✓					1-6	LR
Basile et al. (2017)	✓	✓						3-6	SVM
Castro et al. (2017)	✓	✓						2-7	NB, LR, SVM, RF
Ciobanu et al. (2017)	✓	✓						1-6	SVM
Çöltekin and Rama (2017)	✓	✓	✓					1-7+	SVM
Markov et al. (2017)	✓	✓	✓					3-7	SVM, NB
Martinc et al. (2017)	✓	✓						4	SVM, LR, RF, ...
Medvedeva et al. (2017)	✓	✓	✓	✓				1-6	SVM, CBOW (cf. §6.9)
Mendoza and Mendelsohn (2017)	✓	✓						1-6	SVM
Pla and Hurtado (2017)	✓	✓	✓	✓				1-6	SVM
Williams and Dagli (2017)	✓	✓	✓	✓	✓			1-5	MIRA (cf. §6.8)

Table 8: List of articles where word and character n-grams have been used as features. The numbered columns indicate the length of the word n-grams and char-column the length of character n-grams used. The method column indicates the method used. The relevant section numbers are mentioned in parentheses.

Syntax and part-of-speech ("POS") tags Alex (2005) evaluated methods for detecting foreign language inclusions and experimented with a Conditional Markov Model ("CMM") tagger, which had performed well on Named Entity Recognition ("NER"). Alex (2005) was able to produce the best results by incorporating her own English inclusion classifier's decision as a feature for the tagger, and not using the taggers POS tags. Romsdorfer and Pfister (2007) used syntactic parsers together with dictionaries and morpheme lexicons. Lui and Cook (2013) used n-grams composed of POS tags and function words. Piergallini et al. (2016b) used labels from a NER system, cluster prefixes, and Brown clusters (Brown et al., 1992). Adouane and Dobnik (2017) used POS tag n-grams from one to three and Bestgen (2017) from one to five, and Martinc et al. (2017) used POS tag trigrams with TF-IDF weighting. Schulz and Keller (2016), Basile et al. (2017), van der Lee and Bosch (2017), and Simaki et al. (2017) have also recently used POS tags. Franco-Salvador et al. (2015a) used

POS tags with emotion-labeled graphs in Spanish variety identification. In emotion-labeled graphs, each POS-tag was connected to one or more emotion nodes if a relationship between the original word and the emotion was found from the Spanish Emotion Lexicon. They also used POS-tags with IG-WP. Elfardy et al. (2014) used the MADAMIRA tool for morphological analysis disambiguation. The polySVOX text analysis module described by Romsdorfer and Pfister (2007) uses two-level rules and morpheme lexicons on sub-word level and separate definite clause grammars (DCGs) on word, sentence, and paragraph levels. The language of sub-word units, words, sentences, and paragraphs in multilingual documents is identified at the same time as performing syntactic analysis for the document. Noh et al. (2009) converted sentences into POS-tag patterns using a word-POS dictionary for Malay. The POS-tag patterns were then used by a neural network to indicate whether the sentences were written in Malay or not. Laboreiro et al. (2013) used Jspell to detect differences in the grammar of Portuguese variants. Bekavac et al. (2014) used a syntactic grammar to recognize verb-*da*-verb constructions, which are characteristic of the Serbian language. The syntactic grammar was used together with several morphological grammars to distinguish between Croatian and Serbian.

Languages identified for surrounding words in word-level LI Marcadet et al. (2005) used the weighted LI scores of the words to the left and right of the word to be classified. Rosner and Farrugia (2007) used language labels within an HMM. Akhil and Abhishek (2014) used the language labels of other words in the same sentence to determine the language of the ambiguous word. The languages of the other words had been determined by the positive Decision Rules (Section 6.1), using dictionaries of unique words when possible. Das and Gambäck (2013, 2014) used the language tags of the previous three words with an SVM. Mukherjee et al. (2014) used language labels of surrounding words with NB. King et al. (2015) used the language probabilities of the previous word to determining weights for languages. King et al. (2014b) used unigram, bigram and trigram language label transition probabilities. Papalexakis et al. (2014) used the language labels for the two previous words as well as knowledge of whether code-switching had already been detected or not. Raj and Karfa (2014) used the language label of the previous word to determine the language of an ambiguous word. Sinha and Srinivasa (2014) also used the language label of the previous word. Chanda et al. (2016b) used the language identifications of 2–4 surrounding words for post-identification correction in word-level LI. Samih and Maier (2016) used language labels with a CRF. Dongen (2017) used language labels of the current and two previous words in code-switching point prediction. Their predictive strength was lower than the count of code-switches, but better than the length or position of the word. All of the features were used together with NB, DT and SVM. Guzmán et al. (2017) used language label bigrams with an HMM. Elfardy and Diab (2013) used the word-level language labels obtained with the approach of Elfardy et al. (2013) on sentence-level dialect identification.

5.7 Feature Smoothing

Feature smoothing is required in order to handle the cases where not all features f_i in a test document have been attested in the training corpora. Thus, it is used especially when the count of features is high, or when the amount of training data is low. Smoothing is usually handled as part of the method, and not pre-calculated into the language models. Most of

the smoothing methods evaluated by Chen and Goodman (1999) have been used in LI, and we follow the order of methods in that article.

Additive smoothing (Laplace, Lidstone) In Laplace smoothing, an extra number of occurrences is added to every possible feature in the language model. Dunning (1994) used Laplace's sample size correction (add-one smoothing) with the product of Markovian probabilities. Adams and Resnik (1997) experimented with additive smoothing of 0.5, and noted that it was almost as good as Good-Turing smoothing. Chen and Goodman (1999) calculate the values for each n-gram as:

$$v_{C_g}(f) = \frac{c(C_g, f) + \lambda}{l_{C_g^n} + |U(C_g^n)|\lambda} \tag{11}$$

where $v_{C_g}(f)$ is the probability estimate of n-gram f in the model and $c(C_g, f)$ its frequency in the training corpus. $l_{C_g^n}$ is the total number of n-grams of length n and $|U(C_g^n)|$ the number of distinct n-grams in the training corpus. λ is the Lidstone smoothing parameter. When using Laplace smoothing, λ is equal to 1 and with Lidstone smoothing, the λ is usually set to a value between 0 and 1.

The penalty values used by Jauhiainen et al. (2016) with the HeLI method function as a form of additive smoothing. Vatanen et al. (2010) evaluated additive, Katz, absolute discounting, and Kneser-Ney smoothing methods. Additive smoothing produced the least accurate results of the four methods. Cann (2015) and Franco-Penya and Sanchez (2016) evaluated NB with several different Lidstone smoothing values. Cianflone and Kosseim (2016) used additive smoothing with character n-grams as a baseline classifier, which they were unable to beat with Convolutional Neural Networks ("CNNs").

Good-Turing Discounting Adams and Resnik (1997) used Good-Turing smoothing with the product of Markovian probabilities. Chen and Goodman (1999) define the Good-Turing smoothed count $c_{GT}(C, f)$ as:

$$c_{GT}(C_g, f) = (c(C_g, f) + 1)\frac{r_{c(C_g,f)+1}}{r_{c(C_g,f)}} \tag{12}$$

where $r_{c(C_g,f)}$ is the number of features occurring exactly $c(C_g, f)$ times in the corpus C_g. Lately Good-Turing smoothing has been used by Gamallo et al. (2016) and Giwa (2016).

Jelinek-Mercer Rehůřek and Kolkus (2009) used Jelinek-Mercer smoothing correction over the relative frequencies of words, calculated as follows:

$$v_{C_g}(f) = \lambda\frac{c(C, f)}{l_{C^F}} + (1 - \lambda)\frac{c(C_g, f)}{l_{C_g^F}} \tag{13}$$

where λ is a smoothing parameter, which is usually some small value like 0.1. Mendizabal et al. (2014) used character 1–8 grams with Jelinek-Mercer smoothing. Their language identifier using character 5-grams achieved 3rd place (out of 12) in the TweetLID shared task constrained track.

Katz Ramisch (2008) and Vatanen et al. (2010) used the Katz back-off smoothing (Katz, 1987) from the SRILM toolkit, with perplexity. Katz smoothing is an extension of Good-Turing discounting. The probability mass left over from the discounted n-grams is then distributed over unseen n-grams via a smoothing factor. In the smoothing evaluations by Vatanen et al. (2010), Katz smoothing performed almost as well as absolute discounting, which produced the best results. Giwa and Davel (2013) evaluated Witten-Bell, Katz, and absolute discounting smoothing methods. Witten-Bell got 87.7%, Katz 87.5%, and absolute discounting 87.4% accuracy with character 4-grams.

Prediction by Partial Matching (PPM/Witten-Bell) Teahan (2000) used the PPM-C algorithm for LI. PPM-C is basically a product of Markovian probabilities with an escape scheme. If an unseen context is encountered for the character being processed, the escape probability is used together with a lower-order model probability. In PPM-C, the escape probability is the sum of the seen contexts in the language model. PPM-C was lately used by Adouane et al. (2016d). The PPM-D+ algorithm was used by Celikel (2005). Bergsma et al. (2012) and McNamee (2016) used a PPM-A variant. Yamaguchi and Tanaka-Ishii (2012) also used PPM. The language identifier of Yamaguchi and Tanaka-Ishii (2012) obtained 91.4% accuracy when classifying 100 character texts between 277 languages. Jaech et al. (2016a) used Witten-Bell smoothing with perplexity.

Herman et al. (2016) used a Chunk-Based Language Model ("CBLM"), which is similar to PPM models.

Absolute discounting Vatanen et al. (2010) used several smoothing techniques with Markovian probabilities. Absolute discounting from the VariKN toolkit performed the best. Vatanen et al. (2010) define the smoothing as follows: a constant D is subtracted from the counts $c(C_g, u_{i-n+1,\ldots,i})$ of all observed n-grams $u_{i-n+1,\ldots,i}$ and the held-out probability mass is distributed between the unseen n-grams in relation to the probabilities of lower order n-grams $P_g(u_i|u_{i-n+2,\ldots,i-1})$, as follows:

$$P_{C_g}(u_i|u_{i-n+1,\ldots,i-1}) = \frac{c(C_g, u_{i-n+1,\ldots,i}) - D}{c(C_g, u_{i-n+1,\ldots,i-1})} + \lambda_{u_{i-n+1,\ldots,i-1}} P_{C_g}(u_i|u_{i-n+2,\ldots,i-1}) \quad (14)$$

where $\lambda_{u_{i-n+1,\ldots,i-1}}$ is a scaling factor that makes the conditional distribution sum to one. Absolute discounting with Markovian probabilities from the VariKN toolkit was later also used by Rodrigues (2012), Maier and Gómez-Rodríguez (2014), and Jauhiainen et al. (2017a).

Kneser-Ney smoothing The original Kneser-Ney smoothing is based on absolute discounting with an added back-off function to lower-order models (Vatanen et al., 2010). Chen and Goodman (1999) introduced a modified version of the Kneser-Ney smoothing using interpolation instead of back-off. Chen and Maison (2003) used the Markovian probabilities with Witten-Bell and modified Kneser-Ney smoothing. Giwa (2016), Balažević et al. (2016), and Rijhwani et al. (2017) also recently used modified Kneser-Ney discounting. Barbaresi (2016) used both original and modified Kneser-Ney smoothings. In the evaluations of Vatanen et al. (2010), Kneser-Ney smoothing fared better than additive, but somewhat worse than the Katz and absolute discounting smoothing. Lately Samih and Maier (2016) also used Kneser-Ney smoothing.

Castro et al. (2016, 2017) evaluated several smoothing techniques with character and word n-grams: Laplace/Lidstone, Witten-Bell, Good-Turing, and Kneser-Ney. In their

evaluations, additive smoothing with 0.1 provided the best results. Good-Turing was not as good as additive smoothing, but better than Witten-Bell and Kneser-Ney smoothing. Witten-Bell proved to be clearly better than Kneser-Ney.

6. Methods

In recent years there has been a tendency towards attempting to combine several different types of features into one classifier or classifier ensemble. Many recent studies use readily available classifier implementations and simply report how well they worked with the feature set used in the context of their study. There are many methods presented in this article that are still not available as out of the box implementations, however. There are many studies which have not been re-evaluated at all, going as far back as Mustonen (1965). Our hope is that this article will inspire new studies and many previously unseen ways of combining features and methods. In the following sections, the reviewed articles are grouped by the methods used for LI.

6.1 Decision Rules

Henrich (1989) used a positive Decision Rules with unique characters and character n-grams, that is, if a unique character or character n-gram was found, the language was identified. The positive Decision Rule (unique features) for the test document M and the training corpus C_g can be formulated as follows:

$$R_{DR^+}(g, M) = \begin{cases} 1 & \text{, if } \exists f \in U(M) : c(C_g, f) > 0 \wedge c(C_j, u) = 0 \wedge g \neq j \\ 0 & \text{, otherwise} \end{cases} \tag{15}$$

where $U(M)$ is the set of unique features in M, C_g is the corpus for language g, and C_j is a corpus of any other language j. Positive decision rules can also be used with non-unique features when the decisions are made in a certain order. For example, Dongen (2017) presents the pseudo code for her dictionary lookup tool, where these kind of decisions are part of an if-then-else statement block. Her (manual) rule-based dictionary lookup tool works better for Dutch–English code-switching detection than the SVM, DT, or CRF methods she experiments with. The positive Decision Rule has also been used recently by Abainia et al. (2016), Chanda et al. (2016a,b), Guellil and Azouaou (2016), Gupta et al. (2016), He et al. (2016), and Adouane and Dobnik (2017).

In the negative Decision Rule, if a character or character combination that was found in M does not exist in a particular language, that language is omitted from further identification. The negative Decision Rule can be expressed as:

$$R_{DR^-}(g, M) = \begin{cases} 0 & \text{, if } \exists f \in U(M) : c(C_g, f) = 0 \\ 1 & \text{, otherwise} \end{cases} \tag{16}$$

where C_g is the corpus for language g. The negative Decision Rule was first used by Giguet (1995) in LI.

Alshutayri et al. (2016) evaluated the JRIP classifier from the Waikato Environment for Knowledge Analysis ("WEKA"). JRIP is an implementation of the propositional rule learner. It was found to be inferior to the SVM, NB and DT algorithms.

In isolation the desicion rules tend not to scale well to larger numbers of languages (or very short test documents), and are thus mostly used in combination with other LI methods or as a Decision Tree.

6.2 Decision Trees

Häkkinen and Tian (2001) were the earliest users of Decision Trees ("DT") in LI. They used DT based on characters and their context without any frequency information. In training the DT, each node is split into child nodes according to an information theoretic optimization criterion. For each node a feature is chosen, which maximizes the information gain at that node. The information gain is calculated for each feature and the feature with the highest gain is selected for the node. In the identification phase, the nodes are traversed until only one language is left (leaf node). Later, Ceylan and Kim (2009), Eskander et al. (2014), and Moodley (2016) have been especially successful in using DTs.

Random Forest (RF) is an ensemble classifier generating many DTs. It has been succesfully used in LI by Jhamtani et al. (2014), Darwish et al. (2014), Ranjan et al. (2016), and Malmasi and Zampieri (2017b,a).

6.3 Simple Scoring

In simple scoring, each feature in the test document is checked against the language model for each language, and languages which contain that feature are given a point, as follows:

$$R_{simple}(g, M) = \sum_{i=1}^{l_{MF}} \begin{cases} 1 & \text{, if } f_i \in dom(O(C_g)) \\ 0 & \text{, otherwise} \end{cases} \tag{17}$$

where f_i is the ith feature found in the test document M. The language scoring the most points is the winner. Simple scoring is still a good alternative when facing an easy problem such as preliminary language group identification. It was recently used for this purpose by Franco-Salvador et al. (2015b) with a basic dictionary. They achieved 99.8% accuracy when identifying between 6 language groups. Kadri and Moussaoui (2013) use a version of simple scoring as a distance measure, assigning a penalty value to features not found in a model. In this version, the language scoring the least amount of points is the winner. Their language identifier obtained 100% success rate with character 4-grams when classifying relatively large documents (from 1 to 3 kilobytes), between 10 languages. Simple scoring was also used lately by Balažević et al. (2016), Selamat and Akosu (2016), and Duvenhage et al. (2017).

6.4 Sum or Average of Values

The sum of values can be expressed as:

$$R_{sum}(g, M) = \sum_{i=1}^{l_{MF}} v_{C_g}(f_i) \tag{18}$$

where f_i is the ith feature found in the test document M, and $v_{C_g}(f_i)$ is the value for the feature in the language model of the language g. The language with the highest score is the winner.

The simplest case of Equation 18 is when the text to be identified contains only one feature. An example of this is Shrestha (2014) who used the frequencies of short words as values in word-level identification. For longer words, he summed up the frequencies of different-sized n-grams found in the word to be identified. Giwa and Davel (2014) first calculated the language corresponding to each graphone. They then summed up the predicted languages, and the language scoring the highest was the winner. When a tie occurred, they used the product of the Markovian graphone n-grams. Their method managed to outperform SVMs in their tests.

Henrich (1989) used the average of all the relative frequencies of the n-grams in the text to be identified. Vogel and Tresner-Kirsch (2012) evaluated several variations of the LIGA algorithm introduced by Tromp and Pechenizkiy (2011). Moodley (2016) and Jauhiainen et al. (2017a) also used LIGA and logLIGA methods. The average or sum of relative frequencies was also used recently by Abainia et al. (2016) and Martadinata et al. (2016).

Ng and Selamat (2009) summed up LFDF values (see Section 5.2), obtaining 99.75% accuracy when classifying document sized texts between four languages using Arabic script. Vitale (1991) calculates the score of the language for the test document M as the average of the probability estimates of the features, as follows:

$$R_{avg}(g, M) = \sum_{i=1}^{l_{M^F}} \frac{v_{C_g}(f_i)}{l_{M^F}} \tag{19}$$

where l_{M^F} is the number of features in the test document M. Brown (2013) summed weighted relative frequencies of character n-grams, and normalized the score by dividing by the length (in characters) of the test document. Taking the average of the terms in the sums does not change the order of the scored languages, but it gives comparable results between different lengths of test documents.

Vega and Bressan (2001a,b) summed up the feature weights and divided them by the number of words in the test document in order to set a threshold to detect unknown languages. Their language identifier obtained 89% precision and 94% recall when classifying documents between five languages. El-Shishiny et al. (2004) used a weighting method combining alphabets, prefixes, suffixes and words. Elfardy and Diab (2012) summed up values from a word trigram ranking, basic dictionary and morphological analyzer lookup. Akhil and Abhishek (2014) summed up language labels of the surrounding words to identify the language of the current word. Bekavac et al. (2014) summed up points awarded by the presence of morphological and syntactic features. Gamallo et al. (2014) used inverse rank positions as values. Ács et al. (2015) computed the sum of keywords weighted with TF-IDF. Fabra-Boluda et al. (2015) summed up the TF-IDF derived probabilities of words.

6.5 Product of Values

The product of values can be expressed as follows:

$$R_{prod}(g, M) = \prod_{i}^{l_{M^F}} v_{C_g}(f_i) \tag{20}$$

where f_i is the ith feature found in test document M, and $v_{C_g}(f_i)$ is the value for the feature in the language model of language g. The language with the highest score is the winner.

Some form of feature smoothing is usually required with the product of values method to avoid multiplying by zero.

Product of relative frequencies Church (1985) was the first to use the product of relative frequencies and it has been widely used ever since; recent examples include Castro et al. (2016, 2017), Hanani et al. (2017), and Jauhiainen et al. (2017a). Some of the authors use a sum of log frequencies rather than a product of frequencies to avoid underflow issues over large numbers of features, but the two methods yield the same relative ordering, with the proviso that the maximum of multiplying numbers between 0 and 1 becomes the minimum of summing their negative logarithms, as can be inferred from:

$$R_{logsum}(g, M) = -\log(R_{prod}(g, M)) = -\log \prod_{i=1}^{l_{MF}} v_{C_g}(f_i) = \sum_{i=1}^{l_{MF}} -\log(v_{C_g}(f_i)) \qquad (21)$$

Naive Bayes (NB) When (multinomial[2]) NB is used in LI, each feature used has a probability to indicate each language. The probabilities of all features found in the test document are multiplied for each language, and the language with the highest probability is selected, as in Equation 20. Theoretically the features are assumed to be independent of each other, but in practice using features that are functionally dependent can improve classification accuracy (Peng and Schuurmans, 2003).

NB implementations have been widely used for LI, usually with a more varied set of features than simple character or word n-grams of the same type and length. The features are typically represented as feature vectors given to a NB classifier. Mukherjee et al. (2014) trained a NB classifier with language labels of surrounding words to help predict the language of ambiguous words first identified using an SVM. The language identifier used by Tan et al. (2014) obtained 99.97% accuracy with 5-grams of characters when classifying sentence-sized texts between six language groups. Goutte et al. (2014) used a probabilistic model similar to NB. Bhattu and Ravi (2015) used NB and naive Bayes EM, which uses the Expectation–Maximization ("EM") algorithm in a semi-supervised setting to improve accuracy. Ljubešić and Kranjcić (2014) used Gaussian naive Bayes ("GNB", i.e. NB with Gaussian estimation over continuous variables) from scikit-learn.

Bayesian Network Classifiers In contrast to NB, in Bayesian networks the features are not assumed to be independent of each other. The network learns the dependencies between features in a training phase. Fabra-Boluda et al. (2015) used a Bayesian Net classifier in two-staged (group first) LI over the open track of the DSL 2015 shared task. Rangel et al. (2017a) similarly evaluated Bayesian Nets, but found them to perform worse than the other 11 algorithms they tested.

Product of Markovian probabilities House and Neuburg (1977) used the product of the Markovian probabilities of character bigrams. The language identifier created by Brown (2013, 2014), "whatlang", obtains 99.2% classification accuracy with smoothing for 65 character test strings, when distinguishing between 1,100 languages. The product of Markovian probabilities has recently also been used by Samih and Maier (2016) and Mendoza and Mendelsohn (2017).

2. To the best of our knowledge, the multivariate Bernoulli version of NB has never been used for LI. See Giwa (2016) for a possible explanation.

HeLI Jauhiainen et al. (2016) use a word-based backoff method called HeLI. Here, each language is represented by several different language models, only one of which is used for each word found in the test document. The language models for each language are: a word-level language model, and one or more models based on character n-grams of order 1–n_{max}. When a word that is not included in the word-level model is encountered in a test document, the method backs off to using character n-grams of the size n_{max}. If there is not even a partial coverage here, the method backs off to lower order n-grams and continues backing off until at least a partial coverage is obtained (potentially all the way to character unigrams). The LI system of Jauhiainen et al. (2016) implementing the HeLI method attained shared first place in the closed track of the DSL 2016 shared task (Malmasi et al., 2016), and was the best method tested by Jauhiainen et al. (2017a) for test documents longer than 30 characters.

6.6 Similarity Measures

Out-of-place method The well-known method of Cavnar and Trenkle (1994) uses overlapping character n-grams of varying sizes based on words. The language models are created by tokenizing the training texts for each language g into words, and then padding each word with spaces, one before and four after. Each padded word is then divided into overlapping character n-grams of sizes 1–5, and the counts of every unique n-gram are calculated over the training corpus. The n-grams are ordered by frequency and k of the most frequent n-grams, $f_1, ..., f_k$, are used as the domain of the language model $O(C_g)$ for the language g. The rank of an n-gram f in language g is determined by the n-gram frequency in the training corpus C_g and denoted $\text{rank}_{C_g}(f)$.

During LI, the test document M is treated in a similar way and a corresponding model $O(M)$ of the K most frequent n-grams is created. Then a distance score is calculated between the model of the test document and each of the language models. The value $v_{C_g}(f)$ is calculated as the difference in ranks between $\text{rank}_{C_g}(f)$ and $\text{rank}_M(f)$ of the n-gram f in the domain $dom(O(M))$ of the model of the test document. If an n-gram is not found in a language model, a special penalty value p is added to the total score of the language for each missing n-gram. The penalty value should be higher than the maximum possible distance between ranks.

$$v_{C_g}(f) = \begin{cases} |\,\text{rank}_M(f) - \text{rank}_{C_g}(f)| & \text{, if } f \in dom(O(C_g)) \\ p & \text{, if } f \notin dom(O(C_g)) \end{cases} \tag{22}$$

The score $R_{CT}(g)$ for each language g is the sum of values, as in Equation 18. The language with the lowest score $R_{CT}(g)$ is selected as the identified language. The method is equivalent to Spearman's measure of disarray (Diaconis and Graham, 1977). The out-of-place method has been widely used in LI literature as a baseline. In the evaluations of Jauhiainen et al. (2017a) for 285 languages, the out-of-place method achieved an F-score of 95% for 35-character test documents. It was the fourth best of the seven evaluated methods for test document lengths over 20 characters.

Local Rank Distance ("LRD") Local Rank Distance (Ionescu, 2013) is a measure of difference between two strings. LRD is calculated by adding together the distances identical units (for example character n-grams) are from each other between the two strings. The dis-

tance is only calculated within a local window of predetermined length. Ionescu and Popescu (2016) and Ionescu and Butnaru (2017) used LRD with a Radial Basis Function ("RBF") kernel (see Section 6.8). For learning they experimented with both Kernel Discriminant Analysis ("KDA") and Kernel Ridge Regression ("KRR"). Franco-Salvador et al. (2017a) also used KDA.

Levenshtein distance Pavan et al. (2010) calculated the Levenshtein distance between the language models and each word in the mystery text. The similary score for each language was the inverse of the sum of the Levenshtein distances. Their language identifier obtained 97.7% precision when classifying texts from two to four words between five languages. Later Guellil and Azouaou (2016) used Levenshtein distance for Algerian dialect identification and Gupta et al. (2016) for query word identification.

Probability difference Botha et al. (2007), Botha and Barnard (2007), Botha (2008), and Botha and Barnard (2012) calculated the difference between probabilities as in Equation 23.

$$R_{diff}(g, M) = \sum_i (v_M(f_i) - v_{C_g}(f_i)) \tag{23}$$

where $v_M(f_i)$ is the probability for the feature f_i in the mystery text and $v_{C_g}(f_i)$ the corresponding probability in the language model of the language g. The language with the lowest score $R(g)$ is selected as the most likely language for the mystery text. Singh (2006, 2010) used the log probability difference and the absolute log probability difference. The log probability difference proved slightly better, obtaining a precision of 94.31% using both character and word n-grams when classifying 100 character texts between 53 language-encoding pairs.

Vectors Depending on the algorithm, it can be easier to view language models as vectors of weights over the target features. In the following methods, each language is represented by one or more feature vectors. Methods where each feature type is represented by only one feature vector are also sometimes referred to as centroid-based (Takçı and Güngör, 2012) or nearest prototype methods. Distance measures are generally applied to all features included in the feature vectors.

Kruengkrai et al. (2005) calculated the squared Euclidean distance between feature vectors. The Squared Euclidean distance can be calculated as:

$$R_{\text{euc}^2}(g, M) = \sum_i (v_M(f_i) - v_{C_g}(f_i))^2 \tag{24}$$

Hamzah (2010) used the simQ similarity measure, which is closely related to the Squared Euclidean distance.

Stensby et al. (2010) investigated the LI of multilingual documents using a Stochastic Learning Weak Estimator ("SLWE") method. In SLWE, the document is processed one word at a time and the language of each word is identified using a feature vector representing the current word as well as the words processed so far. This feature vector includes all possible units from the language models – in their case mixed-order character n-grams from one to four. The vector is updated using the SLWE updating scheme to increase the

probabilities of units found in the current word. The probabilities of units that have been found in previous words, but not in the current one, are on the other hand decreased. After processing each word, the distance of the feature vector to the probability distribution of each language is calculated, and the best-matching language is chosen as the language of the current word. Their language identifier obtained 96.0% accuracy when classifying sentences with ten words between three languages. They used the Euclidean distance as the distance measure as follows:

$$R_{\text{euc}}(g, M) = \sqrt{R_{\text{euc}^2}(g, M)} \tag{25}$$

Tomović and Janičić (2007) compared the use of Euclidean distance with their own similarity functions. Prager (1999) calculated the cosine angle between the feature vector of the test document and the feature vectors acting as language models. This is also called the cosine similarity and is calculated as follows:

$$R_{\cos}(g, M) = \frac{\sum_i v_M(f_i) v_{C_g}(f_i)}{\sqrt{\sum_i v_M(f_i)^2} \sqrt{\sum_i v_{C_g}(f_i)^2}} \tag{26}$$

The method of Prager (1999) was evaluated by Lui et al. (2014a) in the context of LI over multilingual documents. The cosine similarity was used recently by Schaetti (2017). One common trick with cosine similarity is to pre-normalise the feature vectors to unit length (e.g. Brown (2012)), in which case the calculation takes the form of the simple dot product:

$$R_{\text{dotprod}}(g, M) = \sum_i v_M(f_i) v_{C_g}(f_i) \tag{27}$$

Jauhiainen (2010) used chi-squared distance, calculated as follows:

$$R_{\textit{chi-square}}(g, M) = \sum_i \frac{(v_{C_g}(f_i) - v_M(f_i))^2}{v_M(f_i)} \tag{28}$$

Abainia et al. (2016) compared Manhattan, Bhattacharyya, chi-squared, Canberra, Bray Curtis, histogram intersection, correlation distances, and out-of-place distances, and found the out-of-place method to be the most accurate.

Entropy Singh (2006, 2010) used cross-entropy and symmetric cross-entropy. Cross-entropy is calculated as follows, where $v_M(f_i)$ and $v_{C_g}(f_i)$ are the probabilities of the feature f_i in the the test document M and the corpus C_g:

$$R_{\textit{cross-entropy}}(g, M) = \sum_i v_M(f_i) \log v_{C_g}(f_i) \tag{29}$$

Symmetric cross-entropy is calculated as:

$$R_{\textit{sym-cross-entropy}}(g, M) = \sum_i v_M(f_i) \log v_{C_g}(f_i) + v_{C_g}(f_i) \log v_M(f_i) \tag{30}$$

For cross-entropy, distribution M must be smoothed, and for symmetric cross-entropy, both probability distributions must be smoothed. Cross-entropy was used recently by

Hanani et al. (2017). Yamaguchi and Tanaka-Ishii (2012) used a cross-entropy estimating method they call the Mean of Matching Statistics ("MMS"). In MMS every possible suffix of the mystery text $u_i, ..., u_{l_M}$ is compared to the language model of each language and the average of the lengths of the longest possible units in the language model matching the beginning of each suffix is calculated.

Sibun and Reynar (1996) and Baldwin and Lui (2010b) calculated the relative entropy between the language models and the test document, as follows:

$$R_{rel\text{-}entropy}(g, M) = \sum_i v_M(f_i) \log \frac{v_M(f_i)}{v_{C_g}(f_i)} \tag{31}$$

This method is also commonly referred to as Kullback-Leibler ("KL") distance or skew divergence. Jauhiainen (2010) compared relative entropy with the product of the relative frequencies for different-sized character n-grams, and found that relative entropy was only competitive when used with character bigrams. The product of relative frequencies gained clearly higher recall with higher-order n-grams when compared with relative entropy.

Singh (2006, 2010) also used the RE and MRE measures, which are based on relative entropy. The RE measure is calculated as follows:

$$R_{RE}(g, M) = \sum_i v_M(f_i) \frac{\log v_M(f_i)}{\log v_{C_g}(f_i)} \tag{32}$$

MRE is the symmetric version of the same measure. In the tests performed by Singh (2006, 2010), the RE measure with character n-grams outperformed other tested methods obtaining 98.51% precision when classifying 100 character texts between 53 language-encoding pairs.

Logistic Regression (LR) Chen and Maison (2003) used a logistic regression ("LR") model (also commonly referred to as "maximum entropy" within NLP), smoothed with a Gaussian prior. Porta and Sancho (2014) defined LR for character-based features as follows:

$$R_{LR}(g, M) = \frac{1}{Z} \exp \sum_j^{l_{M^T}} \sum_i^{l_{C^F}} v_{C_g}(f_i), \text{ if } \exists f_i \in U(M_j^T) \tag{33}$$

where Z is a normalization factor and l_{M^T} is the word count in the word-tokenized test document. Ács et al. (2015) used an LR classifier and found it to be considerably faster than an SVM, with comparable results. Their LR classifier ranked 6 out of 9 on the closed submission track of the DSL 2015 shared task. Lu and Mohamed (2011) used Adaptive Logistic Regression, which automatically optimizes parameters. In recent years LR has been widely used for LI.

Perplexity Ramisch (2008) was the first to use perplexity for LI, in the manner of a language model. He calculated the perplexity for the test document M as follows:

$$H_g(M) = \frac{1}{l_{M^n}} \sum_i^{l_{M^n}} \log_2 v_{C_g}(f_i) \tag{34}$$

$$R_{perplexity}(g, M) = 2^{H_g(M)} \tag{35}$$

where $v_{C_g}(f_i)$ were the Katz smoothed relative frequencies of word n-grams f_i of the length n. Rodrigues (2012) and Jauhiainen et al. (2017a) evaluated the best performing method used by Vatanen et al. (2010). Character n-gram based perplexity was the best method for extremely short texts in the evaluations of Jauhiainen et al. (2017a), but for longer sequences the methods of Brown (2012) and Jauhiainen (2010) proved to be better. Lately, Gamallo et al. (2017) also used perplexity.

Other similarity measures Rau (1974) used Yule's characteristic K and the Kolmogorov-Smirnov goodness of fit test to categorize languages. Kolmogorov-Smirnov proved to be the better of the two, obtaining 89% recall for 53 characters (one punch card) of text when choosing between two languages. In the goodness of fit test, the ranks of features in the models of the languages and the test document are compared. Martins and Silva (2005) experimented with Jiang and Conrath's (JC) distance (Jiang and Conrath, 1997) and Lin's similarity measure (Lin, 1998), as well as the out-of-place method. They conclude that Lin's similarity measure was consistently the most accurate of the three. JC-distance measure was later evaluated by Singh (2006, 2010), and was outperformed by the RE measure. Ranaivo-Malançon and Ng (2005) and Ranaivo-Malançon (2006) calculated special ratios from the number of trigrams in the language models when compared with the text to be identified. da Silva and Lopes (2006b,a, 2007) used the quadratic discrimination score to create the feature vectors representing the languages and the test document. They then calculated the Mahalanobis distance between the languages and the test document. Their language identifier obtained 98.9% precision when classifying texts of four "screen lines" between 19 languages. Nguyen and Cornips (2016) used odds ratio to identify the language of parts of words when identifying between two languages. Odds ratio for language g when compared with language h for morph f_i is calculated as in Equation 36.

$$R_{odds}(g, f_i) = \log \frac{v_{C_h}(f_i)(1 - v_{C_g}(f_i))}{(1 - v_{C_h}(f_i))v_{C_g}(f_i)} \tag{36}$$

6.7 Discriminant Functions

The differences between languages can be stored in discriminant functions. The functions are then used to map the test document into an n-dimensional space. The distance of the test document to the languages known by the language identifier is calculated, and the nearest language is selected (in the manner of a nearest prototype classifier).

Murthy and Kumar (2006) used multiple linear regression to calculate discriminant functions for two-way LI for Indian languages. Bhargava et al. (2015) compared linear regression, NB, and LR. The precision for the three methods was very similar, with linear regression coming second in terms of precision after LR.

Multiple discriminant analysis was used for LI by Mustonen (1965). He used two functions, the first separated Finnish from English and Swedish, and the second separated English and Swedish from each other. He used Mahalanobis' D^2 as a distance measure. Vinosh Babu and Baskaran (2005) used Multivariate Analysis ("MVA") with Principal Component Analysis ("PCA") for dimensionality reduction and LI. Takçı and Ekinci

(2012) compared discriminant analysis with SVM and NN using characters as features, and concluded that the SVM was the best method.

King and Abney (2013) experimented with the Winnow 2 algorithm (Littlestone, 1987), but the method was outperformed by other methods they tested.

6.8 Support Vector Machines ("SVMs")

With support vector machines ("SVMs"), a binary classifier is learned by learning a separating hyperplane between the two classes of instances which maximizes the margin between them. The simplest way to extend the basic SVM model into a multiclass classifier is via a suite of one-vs-rest classifiers, where the classifier with the highest score determines the language of the test document. One feature of SVMs that has made them particularly popular is their compatibility with kernels, whereby the separating hyperplane can be calculated via a non-linear projection of the original instance space. In the following paragraphs, we list the different kernels that have been used with SVMs for LI.

Linear kernel SVMs For LI with SVMs, the predominant approach has been a simple linear kernel SVM model. The linear kernel model has a weight vector $v_{C_g}(f)$ and the classification of a feature vector $v_M(f)$, representing the test document M, is calculated as follows:

$$R_{\text{svm-lin}}(g, M) = (\sum_i v_M(f_i) v_{C_g}(f_i)) + b \tag{37}$$

where b is a scalar bias term. If $R_{\text{svm-lin}}$ is equal to or greater than zero, M is categorized as g.

The first to use a linear kernel SVM were Kim and Park (2007), and generally speaking, linear-kernel SVMs have been widely used for LI, with great success across a range of shared tasks.

Polynomial kernel SVMs Bar and Dershowitz (2014) were the first to apply polynomial kernel SVMs to LI. With a polynomial kernel $R_{\text{svm-pol}}$ can be calculated as:

$$R_{\text{svm-pol}}(g, M) = ((\sum_i v_M(f_i) v_{C_g}(f_i)) + b)^d \tag{38}$$

where d is the polynomial degree, and a hyperparameter of the model.

Radial Basis Function (RBF) kernel SVMs Another popular kernel is the RBF function, also known as a Gaussian or squared exponential kernel. With an RBF kernel $R_{\text{svm-rbf}}$ is calculated as:

$$R_{\text{svm-rbf}}(g, M) = \exp(-\frac{(\sum_i |v_M(f_i) - v_{C_g}(f_i)|)^2}{2\sigma^2}) \tag{39}$$

where σ is a hyperparameter. Botha et al. (2007) were the first to use an RBF kernel SVM for LI.

Reference	linear	string	RBF	sigmoid	d^n	exp.	pas. aggr.
Bhargava and Kondrak (2010)	***	*	**	**	*		
Takçı and Güngör (2012)	*		*	*			
Giwa and Davel (2013); Giwa (2016)	**		***				
Eldesouki et al. (2016)	***	*					
Hanani et al. (2016)	***		*		*		
Xu et al. (2016)	***		*	*	*		
Alrifai et al. (2017)	**				***	*	
Castro et al. (2017)	**						***
Franco-Salvador et al. (2017a)	*		*		*		

Table 9: References where SVMs have been tested with different kernels. The columns indicate the kernels used. "d^n" stands for polynomial kernel.

Sigmoid kernel SVMs With sigmoid kernel SVMs, also known as hyperbolic tangent SVMs, $R_{\text{svm-sig}}$ can be calculated as:

$$R_{\text{svm-sig}}(g, M) = \tanh((\sum_i v_M(f_i) v_{C_g}(f_i)) + b) \tag{40}$$

Bhargava and Kondrak (2010) were the first to use a sigmoid kernel SVM for LI, followed by Majliš (2012), who found the SVM to perform better than NB, Classification And Regression Tree ("CART"), or the sum of relative frequencies.

Other kernels Other kernels that have been used with SVMs for LI include exponential kernels (Alrifai et al., 2017) and rational kernels Porta (2014). Kruengkrai et al. (2005) were the first to use SVMs for LI, in the form of string kernels using Ukkonen's algorithm. They used same string kernels with Euclidean distance, which did not perform as well as SVM. Castro et al. (2017) compared SVMs with linear and on-line passive–aggressive kernels for LI, and found passive–aggressive kernels to perform better, but both SVMs to be inferior to NB and Log-Likelihood Ratio (sum of log-probabilities). Kim and Park (2007) experimented with the Sequential Minimal Optimization ("SMO") algorithm, but found a simple linear kernel SVM to perform better. Alshutayri et al. (2016) achieved the best results using the SMO algorithm, whereas Lamabam and Chakma (2016) found CRFs to work better than SMO. Alrifai et al. (2017) found that SMO was better than linear, exponential and polynomial kernel SVMs for Arabic tweet gender and dialect prediction.

Table 9 lists articles where SVMs with different kernels have been compared. Goutte et al. (2016) evaluated three different SVM approaches using datasets from different DSL shared tasks. SVM-based approaches were the top performing systems in the 2014 and 2015 shared tasks.

Margin Infused Relaxed Algorithm ("MIRA") Williams and Dagli (2017) used SVMs with the Margin Infused Relaxed Algorithm, which is an incremental version of SVM training. In their evaluation, this method achieved better results than off-the-shelf `langid.py`.

6.9 Neural Networks ("NN")

Batchelder (1992) was the first to use Neural Networks ("NN") for LI, in the form of a simple BackPropagation Neural Network ("BPNN") (Hecht-Nielsen, 1989) with a single layer of hidden units, which is also called a multi-layer perceptron ("MLP") model. She used words as the input features for the neural network. Tian et al. (2002) and Tian and Suontausta (2003) succesfully applied MLP to LI.

Jalam and Teytaud (2001a,b) and Jalam (2003) used radial basis function (RBF) networks for LI. Selamat et al. (2007) were the first to use adaptive resonance learning ("ART") neural networks for LI. Abainia et al. (2016) used Neural Text Categorizer ("NTC": Jo (2008)) as a baseline. NTC is an MLP-like NN using string vectors instead of number vectors.

MacNamara et al. (1998) were the first to use a RNN for LI. They concluded that RNNs are less accurate than the simple sum of logarithms of counts of character bi- or trigrams, possibly due to the relatively modestly-sized dataset they experimented with. Babu and Kumar (2010) compared NNs with the out-of-place method (see sec. 6.6). Their results show that the latter, used with bigrams and trigrams of characters, obtains clearly higher identification accuracy when dealing with test documents shorter than 400 characters.

RNNs were more successfully used later by Chang and Lin (2014) who also incorporated character n-gram features in to the network architecture. Cazamias et al. (2015) were the first to use a Long Short-Term Memory ("LSTM") for LI (Hochreiter and Schmidhuber, 1997), and Bjerva (2016) was the first to use Gated Recurrent Unit networks ("GRUs"), both of which are RNN variants. Bjerva (2016) used byte-level representations of sentences as input for the networks. Recently, Hanani et al. (2016) and Samih et al. (2016) also used LSTMs. Later, GRUs were successfully used for LI by Jurgens et al. (2017) and Kocmi and Bojar (2017). In addition to GRUs, Bjerva (2016) also experimented with deep residual networks ("ResNets") at DSL 2016.

During 2016 and 2017, there was a spike in the use of convolutional neural networks (CNNs) for LI, most successfully by Jaech et al. (2016a) and Jaech et al. (2016b). Recently, Li et al. (2018) combined a CNN with adversarial learning to better generalize to unseen domains, surpassing the results of Lui and Baldwin (2012) based on the same training regime as `langid.py`.

Medvedeva et al. (2017) used CBOW NN, achieving better results over the development set of DSL 2017 than RNN-based neural networks. Franco-Salvador et al. (2017b) used deep averaging networks (DANs) based on word embeddings in language variety identification.

6.10 Other Methods

Simaki et al. (2017) used the decision table majority classifier algorithm from the WEKA toolkit in English variety detection. The bagging algorithm using DTs was the best method they tested (73.86% accuracy), followed closely by the decision table with 73.07% accuracy.

Ueda and Nakagawa (1990) were the first to apply hidden Markov models (HMM) to LI. More recently HMMs have been used by Adouane and Dobnik (2017), Guzmán et al. (2017), and Rijhwani et al. (2017). Binas (2005) generated aggregate Markov models, which resulted in the best results when distinguishing between six languages, obtaining 74% accuracy with text length of ten characters. King et al. (2014b) used an extended Markov

Model ("eMM"), which is essentially a standard HMM with modified emission probabilities. Their eMM used manually optimized weights to combine four n-gram scores (products of relative frequencies) into one n-gram score. Xia et al. (2009) used Markov logic networks (Richardson and Domingos, 2006) to predict the language used in interlinear glossed text examples contained in linguistic papers.

Hayta et al. (2013) evaluated the use of unsupervised Fuzzy C Means algorithm ("FCM") in language identification. The unsupervised algorithm was used on the training data to create document clusters. Each cluster was tagged with the language having the most documents in the cluster. Then in the identification phase, the mystery text was mapped to the closest cluster and identified with its language. A supervised centroid classifier based on cosine similarity obtained clearly better results in their experiments (93% vs. 77% accuracy).

Barbaresi (2016) and Martinc et al. (2017) evaluated the extreme gradient boosting ("XGBoost") method (Chen and Guestrin, 2016). Barbaresi (2016) found that gradient boosting gave better results than RFs, while conversely, Martinc et al. (2017) found that LR gave better results than gradient boosting.

Benedetto et al. (2002) used compression methods for LI, whereby a single test document is added to the training text of each language in turn, and the language with the smallest difference (after compression) between the sizes of the original training text file and the combined training and test document files is selected as the prediction. This has obvious disadvantages in terms of real-time computational cost for prediction, but is closely related to language modeling approaches to LI (with the obvious difference that the language model doesn't need to be retrained multiply for each test document). In terms of compression methods, Haţegan et al. (2009) experimented with Maximal Tree Machines ("MTMs"), and Bush (2014) used LZW-based compression.

Very popular in text categorization and topic modeling, Tratz et al. (2013), Tratz (2014), and Voss et al. (2014) used Latent Dirichlet Allocation ("LDA": Blei et al. (2003)) based features in classifying tweets between Arabic dialects, English, and French. Each tweet was assigned with an LDA topic, which was used as one of the features of an LR classifier.

Poulston et al. (2017) used a Gaussian Process classifier with an RBF kernel in an ensemble with an LR classifier. Their ensemble achieved only ninth place in the "PAN" (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection workshop) Author Profiling language variety shared task (Rangel et al., 2017b) and did not reach the results of the baseline for the task.

Espichán-Linares and Oncevay-Marcos (2017, 2018) used a Passive Aggressive classifier, which proved to be almost as good as the SVMs in their evaluations between five different machine learning algorithms from the same package.

6.11 Ensemble Methods

Ensemble methods are meta-classification methods capable of combining several base classifiers into a combined model via a "meta-classifier" over the outputs of the base classifiers, either explicitly trained or some heuristic. It is a simple and effective approach that is used widely in machine learning to boost results beyond those of the individual base classifiers,

and particularly effective when applied to large numbers of individually uncorrelated base classifiers.

Majority and Plurality Voting Rau (1974) used simple majority voting to combine classifiers using different features and methods. In majority voting, the language of the test document is identified if a majority ($> \frac{1}{2}$) of the classifiers in the ensemble vote for the same language. In plurality voting, the language with most votes is chosen as in the simple scoring method (Equation 17). Some authors also refer to plurality voting as majority voting.

Carter et al. (2013) used majority voting in tweet LI. Giwa and Davel (2014) used majority voting with JSM classifiers. Goutte et al. (2014) and Malmasi and Dras (2015a) used majority voting between SVM classifiers trained with different features. Gupta et al. (2014) used majority voting to combine four classifiers: RF, random tree, SVM, and DT. Doval et al. (2014) and Lui and Baldwin (2014) used majority voting between three off-the-shelf language identifiers. Leidig (2014) used majority voting between perplexity-based and other classifiers. Zamora et al. (2014) used majority voting between three sum of relative frequencies-based classifiers where values were weighted with different weighting schemes. Malmasi and Dras (2015b, 2017), Malmasi and Zampieri (2016, 2017b,a), and Mendoza and Mendelsohn (2017) used plurality voting with SVMs.

Gamallo et al. (2017) used voting between several perplexity-based classifiers with different features at the 2017 DSL shared task. A voting ensemble gave better results on the closed track than a singular word-based perplexity classifier (0.9025 weighted F1-score over 0.9013), but worse results on the open track (0.9016 with ensemble and 0.9065 without).

Highest Probability Ensemble In a highest probability ensemble, the winner is simply the language which is given the highest probability by any of the individual classifiers in the ensemble. You et al. (2008) used Gaussian Mixture Models ("GMM") to give probabilities to the outputs of classifiers using different features. Doval et al. (2014) used higher confidence between two off-the-shelf language identifiers. Goutte et al. (2014) used GMM to transform SVM prediction scores into probabilities. Malmasi and Dras (2015b, 2017) used highest confidence over a range of base SVMs. Malmasi and Dras (2017) used an ensemble composed of low-dimension hash-based classifiers. According to their experiments, hashing provided up to 86% dimensionality reduction without negatively affecting performance. Their probability-based ensemble obtained 89.2% accuracy, while the voting ensemble got 88.7%. Balažević et al. (2016) combined an SVM and a LR classifier.

Mean Probability Rule A mean probability ensemble can be used to combine classifiers that produce probabilities (or other mutually comparable values) for languages. The average of values for each language over the classifier results is used to determine the winner and the results are equal to the sum of values method (Equation 18). Malmasi and Dras (2015b) evaluated several ensemble methods and found that the mean probability ensemble attained better results than plurality voting, median probability, product, highest confidence, or Borda count ensembles.

Median Probability Rule In a median probability ensemble, the medians over the probabilities given by the individual classifiers are calculated for each language. Malmasi and Dras (2015b) and Malmasi and Zampieri (2016) used a median probability rule ensemble over

SVM classifiers. Consistent with the results of Malmasi and Dras (2015b), Malmasi and Zampieri (2016) found that a mean ensemble was better than a median ensemble, attaining 68% accuracy vs. 67% for the median ensemble.

Product Rule A product rule ensemble takes the probabilities for the base classifiers and calculates their product (or sum of the log probabilities), with the effect of penalising any language where there is a particularly low probability from any of the base classifiers. Giwa and Davel (2014) used log probability voting with JSM classifiers. Giwa and Davel (2014) observed a small increase in average accuracy using the product ensemble over a majority voting ensemble.

k-best Ensemble (k-best) In a k-best ensemble, several models are created for each language g by partitioning the corpus C_g into separate samples. The score $R(C_{g_i}, M)$ is calculated for each model. For each language, plurality voting is then applied to the k models with the best scores to predict the language of the test document M. Jalam and Teytaud (2001b) evaluated k-best with $k = 1$ based on several similarity measures. Kerwin (2006) compared $k = 10$ and $k = 50$ and concluded that there was no major difference in accuracy when distinguishing between six languages (100 character test set). Baykan et al. (2008) experimented with k-best classifiers, but they gave clearly worse results than the other classifiers they evaluated. Barman et al. (2014b) used k-best in two phases, first selecting $k_1 = 800$ closest neighbors with simple similarity, and then using $k_2 = 16$ with a more advanced similarity ranking.

Bootstrap Aggregating (Bagging) In bagging, independent samples of the training data are generated by random sampling with replacement, individual classifiers are trained over each such training data sample, and the final classification is determined by plurality voting. Martinc et al. (2017) evaluated the use of bagging with an LR classifier in PAN 2017 language variety identification shared task, however, bagging did not improve the accuracy in the 10-fold cross-validation experiments on the training set. Sierra et al. (2017) used bagging with word convolutional neural networks ("W-CNN"). Simaki et al. (2017) used bagging with DTs in English national variety detection and found DT-based bagging to be the best evaluated method when all 60 different features (a wide selection of formal, POS, lexicon-based, and data-based features) were used, attaining 73.86% accuracy. Simaki et al. (2017) continued the experiments using the ReliefF feature selection algorithm from the WEKA toolkit to select the most efficient features, and achieved 77.32% accuracy over the reduced feature set using a NB classifier.

Rotation Forest Rangel et al. (2017a) evaluated the Rotation Forest meta classifier for DTs. The method randomly splits the used features into a pre-determined number of subsets and then uses PCA for each subset. It obtained 66.6% accuracy, attaining fifth place among the twelve methods evaluated.

Adaptive Boosting (AdaBoost) The AdaBoost algorithm (Freund and Schapire, 1997) examines the performance of the base classifiers on the evaluation set and iteratively boosts the significance of misclassified training instances, with a restart mechanism to avoid local minima. AdaBoost was the best of the five machine learning techniques evaluated by van der Lee and Bosch (2017), faring better than C4.5, NB, RF, and linear SVM.

Rangel et al. (2017a) used the LogitBoost variation of AdaBoost. It obtained 67.0% accuracy, attaining third place among the twelve methods evaluated.

Stacked generalization (Stacking) In stacking, a higher level classifier is explicitly trained on the output of several base classifiers. You et al. (2008) used AdaBoost.ECC and CART to combine classifiers using different features. More recently, Mathur et al. (2017) used LR to combine the results of five RNNs. As an ensemble they produced better results than NB and LR, which were better than the individual RNNs. Also in 2017, Malmasi and Zampieri (2017b,a) used RF to combine several linear SVMs with different features. The system used by Malmasi and Zampieri (2017a) ranked first in the German dialect identification shared task, and the system by Malmasi and Zampieri (2017b) came second (71.65% accuracy) in the Arabic dialect identification shared task.

7. Empirical Evaluation

In the previous two sections, we have alluded to issues of evaluation in LI research to date. In this section, we examine the literature more closely, providing a broad overview of the evaluation metrics that have been used, as well as the experimental settings in which LI research has been evaluated.

7.1 Standardized Evaluation for LI

The most common approach is to treat the task as a document-level classification problem. Given a set of evaluation documents, each having a known correct label from a closed set of labels (often referred to as the "gold-standard"), and a predicted label for each document from the same set, the document-level accuracy is the proportion of documents that are correctly labeled over the entire evaluation collection. This is the most frequently reported metric and conveys the same information as the error rate, which is simply the proportion of documents that are incorrectly labeled (i.e. $1 - \text{accuracy}$).

Authors sometimes provide a per-language breakdown of results. There are two distinct ways in which results are generally summarized per-language: (1) precision, in which documents are grouped according to their predicted language; and (2) recall, in which documents are grouped according to what language they are actually written in. Earlier work has tended to only provide a breakdown based on the correct label (i.e. only reporting per-language recall). This gives us a sense of how likely a document in any given language is to be classified correctly, but does not give an indication of how likely a prediction for a given language is of being correct. Under the monolingual assumption (i.e. each document is written in exactly one language), this is not too much of a problem, as a false negative for one language must also be a false positive for another language, so precision and recall are closely linked. Nonetheless, authors have recently tended to explicitly provide both precision and recall for clarity. It is also common practice to report an F-score F, which is the harmonic mean of precision and recall. The F-score (also sometimes called F1-score or F-measure) was developed in IR to measure the effectiveness of retrieval with respect to a user who attaches different relative importance to precision and recall (van Rijsbergen, 1979). When used as an evaluation metric for classification tasks, it is common to place equal

weight on precision and recall (hence "F1"-score, in reference to the β hyper-parameter, which equally weights precision and recall when $\beta = 1$).

In addition to evaluating performance for each individual language, authors have also sought to convey the relationship between classification errors and specific sets of languages. Errors in LI systems are generally not random; rather, certain sets of languages are much more likely to be confused. The typical method of conveying this information is through the use of a confusion matrix, a tabulation of the distribution of (predicted language, actual language) pairs.

Presenting full confusion matrices becomes problematic as the number of languages considered increases, and as a result has become relatively uncommon in work that covers a broader range of languages. Per-language results are also harder to interpret as the number of languages increases, and so it is common to present only collection-level summary statistics. There are two conventional methods for summarizing across a whole collection: (1) giving each document equal weight; and (2) giving each class (i.e. language) equal weight. (1) is referred to as a micro-average, and (2) as a macro-average. For LI under the monolingual assumption, micro-averaged precision and recall are the same, since each instance of a false positive for one language must also be a false negative for another language. In other words, micro-averaged precision and recall are both simply the collection-level accuracy. On the other hand, macro-averaged precision and recall give equal weight to each language. In datasets where the number of documents per language is the same, this again works out to being the collection-level average. However, LI research has frequently dealt with datasets where there is a substantial skew between classes. In such cases, the collection-level accuracy is strongly biased towards more heavily-represented languages. To address this issue, in work on skewed document collections, authors tend to report both the collection-level accuracy and the macro-averaged precision/recall/F-score, in order to give a more complete picture of the characteristics of the method being studied.

Whereas the notions of macro-averaged precision and recall are clearly defined, there are two possible methods to calculate the macro-averaged F-score. The first is to calculate it as the harmonic mean of the macro-averaged precision and recall, and the second is to calculate it as the arithmetic mean of the per-class F-score.

The comparability of published results is also limited by the variation in size and source of the data used for evaluation. In work to date, authors have used data from a variety of different sources to evaluate the performance of proposed solutions. Typically, data for a number of languages is collected from a single source, and the number of languages considered varies widely. Earlier work tended to focus on a smaller number of Western European languages. Later work has shifted focus to supporting larger numbers of languages simultaneously, with the work of Brown (2014) pushing the upper bound, reporting a language identifier that supports over 1300 languages. The increased size of the language set considered is partly due to the increased availability of language-labeled documents from novel sources such as Wikipedia and Twitter. This supplements existing data from translations of the Universal Declaration of Human Rights, bible translations, as well as parallel texts from MT datasets such as OPUS and SETimes, and European Government data such as JRC-Acquis. These factors have led to a shift away from proprietary datasets such as the ECI multilingual corpus that were commonly used in earlier research. As more languages are considered simultaneously, the accuracy of LI systems decreases. A particularly striking

illustration of this is the evaluation results by Jauhiainen et al. (2017a) for the logLIGA method (Vogel and Tresner-Kirsch, 2012). Vogel and Tresner-Kirsch (2012) report an accuracy of 99.8% over tweets (averaging 80 characters) in six European languages as opposed to the 97.9% from the original LIGA method. The LIGA and logLIGA implementations by Jauhiainen et al. (2017a) have comparable accuracy for six languages, but the accuracy for 285 languages (with 70 character test length) is only slightly over 60% for logLIGA and the original LIGA method is at almost 85%. Many evaluations are not directly comparable as the test sizes, language sets, and hyper-parameters differ. A particularly good example is the method of Cavnar and Trenkle (1994). The original paper reports an accuracy of 99.8% over eight European languages (¿300 bytes test size). Lui and Baldwin (2011) report an accuracy of 68.6% for the method over a dataset of 67 languages (500 byte test size), and Jauhiainen et al. (2017a) report an accuracy of over 90% for 285 languages (25 character test size).

Separate to the question of the number and variety of languages included are issues regarding the quantity of training data used. A number of studies have examined the relationship between LI accuracy and quantity of training data through the use of learning curves. The general finding is that LI accuracy increases with more training data, though there are some authors that report an optimal amount of training data, where adding more training data decreases accuracy thereafter (Ljubešić et al., 2007). Overall, it is not clear whether there is a universal quantity of data that is "enough" for any language, rather this amount appears to be affected by the particular set of languages as well as the domain of the data. The breakdown presented by Baldwin and Lui (2010b) shows that with less than 100KB per language, there are some languages where classification accuracy is near perfect, whereas there are others where it is very poor.

Another aspect that is frequently reported on is how long a sample of text needs to be before its language can be correctly detected. Unsurprisingly, the general consensus is that longer samples are easier to classify correctly. There is a strong interest in classifying short segments of text, as certain applications naturally involve short text documents, such as LI of microblog messages or search engine queries. Another area where LI of texts as short as one word has been investigated is in the context of dealing with documents that contain text in more than one language, where word-level LI has been proposed as a possible solution (see Section 10.6). These outstanding challenges have led to research focused specifically on LI of shorter segments of text, which we discuss in more detail in Section 10.7.

From a practical perspective, knowing the rate at which a LI system can process and classify documents is useful as it allows a practitioner to predict the time required to process a document collection given certain computational resources. However, so many factors influence the rate at which documents are processed that comparison of absolute values across publications is largely meaningless. Instead, it is more valuable to consider publications that compare multiple systems under controlled conditions (same computer hardware, same evaluation data, etc.). The most common observations are that classification times between different algorithms can differ by orders of magnitude, and that the fastest methods are not always the most accurate. Beyond that, the diversity of systems tested and the variety in the test data make it difficult to draw further conclusions about the relative speed of algorithms.

Where explicit feature selection is used, the number of features retained is a parameter of interest, as it affects both the memory requirements of the LI system and its classification rate. In general, a smaller feature set results in a faster and more lightweight identifier. Relatively few authors give specific details of the relationship between the number of features selected and accuracy. A potential reason for this is that the improvement in accuracy plateaus with increasing feature count, though the exact number of features required varies substantially with the method and the data used. At the lower end of the scale, Cavnar and Trenkle (1994) report that 300–400 features per language is sufficient. Conversely Jauhiainen et al. (2017a) found that, for the same method, the best results for the evaluation set were attained with 20,000 features per language.

7.2 Corpora Used for LI Evaluation

As discussed in Section 7.1, the objective comparison of different methods for LI is difficult due to the variation in the data that different authors have used to evaluate LI methods. Baldwin and Lui (2010b) emphasize this by demonstrating how the performance of a system can vary according to the data used for evaluation. This implies that comparisons of results reported by different authors may not be meaningful, as a strong result in one paper may not translate into a strong result on the dataset used in a different paper. In other areas of research, authors have proposed standardized corpora to allow for the objective comparison of different methods.

Some authors have released datasets to accompany their work, to allow for direct replication of their experiments and encourage comparison and standardization. Table 10 lists a number of datasets that have been released to accompany specific LI publications. In this list, we only include corpora that were prepared specifically for LI research, and that include the full text of documents. Corpora of language-labelled Twitter messages that only provide document identifiers are also available, but reproducing the full original corpus is always an issue as the original Twitter messages are deleted or otherwise made unavailable.

One challenge in standardizing datasets for LI is that the codes used to label languages are not fully standardized, and a large proportion of labeling systems only cover a minor portion of the languages used in the world today (Constable and Simons, 2000). Xia et al. (2010) discuss this problem in detail, listing different language code sets, as well as the internal structure exhibited by some of the code sets. Some standards consider certain groups of "languages" as varieties of a single macro-language, whereas others consider them to be discrete languages. An example of this is found in South Slavic languages, where some language code sets refer to Serbo-Croatian, whereas others make distinctions between Bosnian, Serbian and Croatian (Tiedemann and Ljubešić, 2012). The unclear boundaries between such languages make it difficult to build a reference corpus of documents for each language, or to compare language-specific results across datasets.

Another challenge in standardizing datasets for LI is the great deal of variation that can exist between data in the same language. We examine this in greater detail in Section 10.2, where we discuss how the same language can use a number of different orthographies, can be digitized using a number of different encodings, and may also exist in transliterated forms. The issue of variation within a language complicates the development of standardized datasets, due to challenges in determining which variants of a language should be

Reference	Type	Source
Baldwin and Lui (2010b)	Multilingual (81)	Government Documents, News Texts, Wikipedia
https://github.com/varh1i/language_detection/tree/master/src/main/resources/naacl2010-langid		
Baldwin and Lui (2010a)	Multilingual (74)	Wikipedia (synthetic multilingual docs)
http://people.eng.unimelb.edu.au/tbaldwin/etc/altw2010-langid.tgz		
Vatanen et al. (2010)	Multiglingual (281)	Universal Declaration of Human Rights (UDHR)
http://research.ics.aalto.fi/cog/data/udhr/		
Tromp and Pechenizkiy (2011)	Multilingual (6)	Twitter
http://www.win.tue.nl/~mpechen/projects/smm/LIGA_Benelearn11_dataset.zip		
Zaidan and Callison-Burch (2011)	Arabic dialects (5)	Arabic Online Commentary (AOC)
https://github.com/sjeblee/AOC		
Lui and Baldwin (2011)	Multilingual	Various
http://people.eng.unimelb.edu.au/tbaldwin/etc/ijcnlp2011-langid.tgz		
Majliš (2012)	Multilingual (124)	Wikipedia (YALI)
http://ufal.mff.cuni.cz/tools/yali		
Tiedemann and Ljubešić (2012)	Multilingual (10)	SETimes (News Texts)
http://www.nljubesic.net/resources/corpora/setimes/		
Brown (2013)	Multilingual (970/1279)	Bible Translations, Wikipedia
http://sourceforge.net/projects/la-strings/files/Language-Data/		
King and Abney (2013)	Multilingual (30)	Web Crawl
http://www-personal.umich.edu/~benking/resources/mixed-language-annotations-release-v1.0.tgz		
Lui and Baldwin (2014)	Multilingual (65)	Twitter
http://people.eng.unimelb.edu.au/tbaldwin/data/lasm2014-twituser-v1.tgz		
Lui et al. (2014a)	Multilingual (156)	WikipediaMulti
http://people.eng.unimelb.edu.au/tbaldwin/etc/wikipedia-multi-v6.tgz		
Tan et al. (2014)	Multilingual (22)	News Texts
http://ttg.uni-saarland.de/resources/DSLCC/		
Chanda et al. (2016a)	Code-switching (2)	Facebook chats
https://github.com/ArunavhaChanda/Facebook-Code-Mixed-Corpus		
Blodgett et al. (2017)	Multilingual (70)	Twitter70
http://slanglab.cs.umass.edu/TwitterLangID/		

Table 10: Published LI Datasets

included. Since we have seen that the performance of LI systems can vary per-domain (Baldwin and Lui, 2010b), that LI research is often motivated by target applications (see Section 8), and that domain-specific information can be used to improve accuracy (see Section 10.9), it is often unsound to use a generic LI dataset to develop a language identifier for a particular domain.

A third challenge in standardizing datasets for LI is the cost of obtaining correctly-labeled data. Manual labeling of data is usually prohibitively expensive, as it requires access to native speakers of all languages that the dataset aims to include. Large quantities of raw text data are available from sources such as web crawls or Wikipedia, but this data is frequently mislabeled (e.g. most non-English Wikipedias still include some English-language documents). In constructing corpora from such resources, it is common to use some form of automatic LI, but this makes such corpora unsuitable for evaluation purposes as they are biased towards documents that can be correctly identified by automatic systems (Lui and Baldwin, 2014). Future work in this area could investigate other means of ensuring correct gold-standard labels while minimizing the annotation cost.

Despite these challenges, standardized datasets are critical for replicable and comparable research in LI. Where a subset of data is used from a larger collection, researchers should

include details of the specific subset, including any breakdown into training and test data, or partitions for cross-validation. Where data from a new source is used, justification should be given for its inclusion, as well as some means for other researchers to replicate experiments on the same dataset.

7.3 LI Shared Tasks

To address specific sub-problems in LI, a number of shared tasks have been organized on problems such as LI in multilingual documents (Baldwin and Lui, 2010a), code-switched data (Solorio et al., 2014), discriminating between closely related languages (Zampieri et al., 2014), and dialect and language variety identification in various languages (Grouin et al., 2011; Zampieri et al., 2017; Rangel et al., 2017b; Ali et al., 2017). Shared tasks are important for LI because they provide datasets and standardized evaluation methods that serve as benchmarks for the LI community. We summarize all LI shared tasks organized to date in Table 11.

Generally, datasets for shared tasks have been made publicly available after the conclusion of the task, and are a good source of standardized evaluation data. However, the shared tasks to date have tended to target specific sub-problems in LI, and no general, broad-coverage LI datasets have been compiled. Widespread interest in LI over closely-related languages has resulted in a number of shared tasks that specifically tackle the issue. Some tasks have focused on varieties of a specific language. For example, the DEFT2010 shared task (Grouin et al., 2011) examined varieties of French, requiring participants to classify French documents with respect to their geographical source, in addition to the decade in which they were published. Another example is the Arabic Dialect Identification ("ADI") shared task at the VarDial workshop (Malmasi, 2017; Zampieri et al., 2017), and the Arabic Multi-Genre Broadcast ("MGB") Challenge (Ali et al., 2017).

Two shared tasks focused on a narrow group of languages using Twitter data. The first was TweetLID, a shared task on LI of Twitter messages according to six languages in common use in Spain, namely: Spanish, Portuguese, Catalan, English, Galician, and Basque (in order of the number of documents in the dataset) (Zubiaga et al., 2014, 2016). The organizers provided almost 35,000 Twitter messages, and in addition to the six monolingual tags, supported four additional categories: undetermined, multilingual (i.e. the message contains more than one language, without requiring the system to specify the component languages), ambiguous (i.e. the message is ambiguous between two or more of the six target languages), and other (i.e. the message is in a language other than the six target languages). The second shared task was the PAN lab on authorship profiling 2017 (Rangel et al., 2017b). The PAN lab on authorship profiling is held annually and historically has focused on age, gender, and personality traits prediction in social media. In 2017 the competition introduced the inclusion of language varieties and dialects of Arabic, English, Spanish, and Portuguese,

More ambitiously, the four editions of the Discriminating between Similar Languages (DSL) (Zampieri et al., 2014, 2015b; Malmasi et al., 2016; Zampieri et al., 2017) shared tasks required participants to discriminate between a set of languages in several language groups, each consisting of highly-similar languages or national varieties of that language. The dataset, entitled DSL Corpus Collection ("DSLCC") (Tan et al., 2014), and the languages included are summarized in Table 12. Historically the best-performing systems

Year – Title	Reference
2010 – DÉfi Fouille de Texte (DEFT) https://deft.limsi.fr	Grouin et al. (2011)
2010 – Australasian Language Technology Workshop http://www.alta.asn.au	Baldwin and Lui (2010a)
2014 – Twitter Language Identification Workshop at SEPLN 2014 http://komunitatea.elhuyar.org/tweetlid/?lang=en_us	Zubiaga et al. (2014)
2014 – Computational Approaches to Code Switching http://emnlp2014.org/workshops/CodeSwitch/call.html	Solorio et al. (2014)
2014 – First DSL Shared Task at VarDial http://corporavm.uni-koeln.de/vardial/sharedtask.html	Zampieri et al. (2014)
2015 – Second DSL Shared Task at VarDial http://ttg.uni-saarland.de/lt4vardial2015/dsl.html	Zampieri et al. (2015b)
2016 – First Arabic Dialect Identification (ADI) at VarDial http://ttg.uni-saarland.de/vardial2016/dsl2016.html	Malmasi et al. (2016)
2016 – Third DSL Shared Task at VarDial http://ttg.uni-saarland.de/vardial2016/dsl2016.html	Malmasi et al. (2016)
2017 – Second Arabic Dialect Identification (ADI) at VarDial http://ttg.uni-saarland.de/vardial2017/sharedtask2017.html	Zampieri et al. (2017)
2017 – Fourth DSL Shared Task at VarDial http://ttg.uni-saarland.de/vardial2017/sharedtask2017.html	Zampieri et al. (2017)
2017 – First German Dialect Identification (ADI) at VarDial http://ttg.uni-saarland.de/vardial2017/sharedtask2017.html	Zampieri et al. (2017)
2017 – PAN lab on Author Profiling http://pan.webis.de/clef17/pan17-web/author-profiling.html	Rangel et al. (2017b)
2017 – Arabic Multi-Genre Broadcast (MGB) Challenge http://www.mgb-challenge.org/arabic.html	Ali et al. (2017)

Table 11: List of LI shared tasks.

(Goutte et al., 2014; Lui et al., 2014b; Bestgen, 2017) have approached the task via hierarchical classification, first predicting the language group, then the language within that group.

8. Application Areas

There are various reasons to investigate LI. Studies in LI approach the task from different perspectives, and with different motivations and application goals in mind. In this section, we briefly summarize what these motivations are, and how their specific needs differ.

The oldest motivation for automatic LI is perhaps in conjunction with translation (Beesley, 1988). Automatic LI is used as a pre-processing step to determine what translation model to apply to an input text, whether it be by routing to a specific human translator or by applying MT. Such a use case is still very common, and can be seen in the Google Chrome web browser,[3] where an built-in LI module is used to offer MT services to the user

3. http://www.google.com/chrome

Language/Variety	v1.0 (2014)	v2.0/2.1 (2015)	v3.0 (2016)	v4.0 (2017)
Bosnian	✓	✓	✓	✓
Croatian	✓	✓	✓	✓
Serbian	✓	✓	✓	✓
Czech	✓	✓		
Slovak	✓	✓		
Indonesian	✓	✓	✓	✓
Malay	✓	✓	✓	✓
Brazilian Portuguese	✓	✓	✓	✓
European Portuguese	✓	✓	✓	✓
Macanese Portuguese		✓		
Argentine Spanish	✓	✓	✓	✓
Castilian Spanish	✓	✓	✓	✓
Mexican Spanish		✓	✓	
Peruvian Spanish				✓
Bulgarian		✓		
Macedonian		✓		
Canadian French			✓	✓
Hexagonal French			✓	✓
American English	✓			
British English	✓			
Persian			✓	
Dari			✓	

Table 12: DSLCC: the languages included in each version of the corpus collection, grouped by language similarity.

when the detected language of the web page being visited differs from the user's language settings.

NLP components such as POS taggers and parsers tend to make a strong assumption that the input text is monolingual in a given language. Similarly to the translation case, LI can play an obvious role in routing documents written in different languages to NLP components tailored to those languages. More subtle is the case of documents with mixed multilingual content, the most commonly-occurring instance of which is foreign inclusion, where a document is predominantly in a single language (e.g. German or Japanese) but is interspersed with words and phrases (often technical terms) from a language such as English. For example, Alex et al. (2007) found that around 6% of word tokens in German text sourced from the Internet are English inclusions. In the context of POS tagging, one strategy for dealing with inclusions is to have a dedicated POS for all foreign words, and force the POS tagger to perform both foreign inclusion detection and POS tag these words in the target language; this is the approach taken in the Penn POS tagset, for example (Marcus et al., 1993). An alternative strategy is to have an explicit foreign inclusion detection pre-processor, and some special handling of foreign inclusions. For example, in the context of German parsing, Alex et al. (2007) used foreign inclusion predictions to restrict the set of (German) POS tags used to form a parse tree, and found that this approach substantially improved parser accuracy.

Another commonly-mentioned use case is for multilingual document storage and retrieval. A document retrieval system (such as, but not limited to, a web search engine) may

be required to index documents in multiple languages. In such a setting, it is common to apply LI at two points: (1) to the documents being indexed; and (2) to the queries being executed on the collection. Simple keyword matching techniques can be problematic in text-based document retrieval, because the same word can be valid in multiple languages. A classic example of such words (known as "false friends") includes *gift*, which in German means "poison". Performing LI on both the document and the query helps to avoid confusion between such terms, by taking advantage of the context in which it appears in order to infer the language. This has resulted in specific work in LI of web pages, as well as search engine queries. Roy et al. (2013) and Sequeira et al. (2015) give overviews of shared tasks specifically concentrating on language labeling of individual search query words. Having said this, in many cases, the search query itself does a sufficiently good job of selecting documents in a particular language, and overt LI is often not performed in mixed multilingual search contexts.

Automatic LI has also been used to facilitate linguistic and other text-based research. Suzuki et al. (2002) report that their motivation for developing a language identifier was "to find out how many web pages are written in a particular language". Automatic LI has been used in constructing web-based corpora. The Crúbadán project (Scannell, 2007) and the Finno-Ugric Languages and the Internet project (Jauhiainen et al., 2015a) make use of automated LI techniques to gather linguistic resources for under-resourced languages. Similarly, the Online Database of INterlinear text ("ODIN": Lewis and Xia (2010)) uses automated LI as one of the steps in collecting interlinear glossed text from the web for purposes of linguistic search and bootstrapping NLP tools.

One challenge in collecting linguistic resources from the web is that documents can be multilingual (i.e. contain text in more than one language). This is problematic for standard LI methods, which assume that a document is written in a single language, and has prompted research into segmenting text by language, as well as word-level LI, to enable extraction of linguistic resources from multilingual documents. A number of LI shared tasks discussed in detail in Section 7.3 included data from social media. Examples are the TweetLID shared task on tweet LI held at SEPLN 2014 (Zubiaga et al., 2014, 2016), the data sets used in the first and second shared tasks on LI in code-switched data which were partially taken from Twitter (Solorio et al., 2014; Molina et al., 2016), and the third edition of the DSL shared task which contained two out-of-domain test sets consisting of tweets (Malmasi et al., 2016). The 5th edition of the PAN at CLEF author profiling task included language variety identification for tweets (Rangel et al., 2017b). There has also been research on identifying the language of private messages between eBay users (Mayer, 2012), presumably as a filtering step prior to more in-depth data analysis.

9. Off-the-Shelf Language Identifiers

An "off-the-shelf" language identifier is software that is distributed with pre-trained models for a number of languages, so that a user is not required to provide training data before using the system. Such a setup is highly attractive to many end-users of automatic LI whose main interest is in utilizing the output of a language identifier rather than implementing and developing the technique. To this end, a number of off-the-shelf language identifiers have been released over time. Many authors have evaluated these off-the-shelf identifiers,

including a recent evaluation involving 13 language identifiers which was carried out by Pawelka and Jürgens (2015). In this section, we provide a brief summary of open-source or otherwise free systems that are available, as well as the key characteristics of each system. We have also included dates of when the software has been last updated as of October 2018.

`TextCat` is the most well-known Perl implementation of the out-of-place method, it lists models for 76 languages in its off-the-shelf configuration;[4] the program is not actively maintained. `TextCat` is not the only example of an off-the-shelf implementation of the out-of-place method: other implementations include `libtextcat` with 76 language models,[5] `JTCL` with 15 languages,[6] and `mguesser` with 104 models for different language-encoding pairs.[7] The main issue addressed by later implementations is classification speed: `TextCat` is implemented in Perl and is not optimized for speed, whereas implementations such as `libtextcat` and `mguesser` have been specifically written to be fast and efficient. `whatlang-rs` uses an algorithm based on character trigrams and refers the user to the Cavnar and Trenkle (1994) article. It comes pre-trained with 83 languages.[8]

`ChromeCLD` is the language identifier embedded in the Google Chrome web browser.[9] It uses a NB classifier, and script-specific classification strategies. `ChromeCLD` assumes that all the input is in UTF-8, and assigns the responsibility of encoding detection and transcoding to the user. `ChromeCLD` uses Unicode information to determine the script of the input. `ChromeCLD` also implements a number of pre-processing heuristics to help boost performance on its target domain (web pages), such as stripping character sequences like `.jpg`. The standard implementation supports 83 languages, and an extended model is also available, that supports 160 languages.[10]

`LangDetect` is a Java library that implements a language identifier based on a NB classifier trained over character n-grams. The software comes with pre-trained models for 53 languages, using data from Wikipedia.[11] `LangDetect` makes use of a range of normalization heuristics to improve the performance on particular languages, including: (1) clustering of Chinese/Japanese/Korean characters to reduce sparseness; (2) removal of "language-independent" characters, and other text normalization; and (3) normalization of Arabic characters.

`langid.py` is a Python implementation of the method described by Lui and Baldwin (2011), which exploits training data for the same language across multiple different sources of text to identify sequences of characters that are strongly predictive of a given language, regardless of the source of the text. This feature set is combined with a NB classifier, and is distributed with a pre-trained model for 97 languages prepared using data from 5 different text sources.[12] Lui and Baldwin (2012) provide an empirical comparison of `langid.py` to `TextCat`, `LangDetect` and `ChromeCLD` and find that it compares favorably both in

4. `http://odur.let.rug.nl/~vannoord/TextCat/`
5. `https://software.wise-guys.nl/libtextcat/` (not updated since 2003)
6. `http://textcat.sourceforge.net` (not updated since first release)
7. `http://www.mnogosearch.org/guesser/` (not updated since 2008)
8. `https://github.com/greyblake/whatlang-rs` (last updated June 2018)
9. `http://www.google.com/chrome`
10. `https://github.com/CLD2Owners/cld2` (last updated on August 2015)
11. `https://github.com/shuyo/language-detection` (last updated on March 2014)
12. `https://github.com/saffsd/langid.py` (last updated on July 2017)

terms of accuracy and classification speed. There are also implementations of the classifier component (but not the training portion) of `langid.py` in Java, [13] C,[14] and JavaScript.[15]

`whatlang` (Brown, 2013) uses a vector-space model with per-feature weighting on character n-gram sequences. One particular feature of `whatlang` is that it uses discriminative training in selecting features, i.e. it specifically makes use of features that are strong evidence *against* a particular language, which is generally not captured by NB models. Another feature of `whatlang` is that it uses inter-string smoothing to exploit sentence-level locality in making language predictions, under the assumption that adjacent sentences are likely to be in the same language. Brown (2013) reports that this substantially improves the accuracy of the identifier. Another distinguishing feature of `whatlang` is that it comes pre-trained with data for 1400 languages, which is the highest number by a large margin of any off-the-shelf system.[16]

`whatthelang` is a recent language identifier written in Python, which utilizes the FastText NN-based text classification algorithm. It supports 176 languages.[17]

`YALI` implements an off-the-shelf classifier trained using Wikipedia data, covering 122 languages.[18] Although not described as such, the actual classification algorithm used is a linear model, and is thus closely related to both NB and a cosine-based vector space model.

In addition to the above-mentioned general-purpose language identifiers, there have also been efforts to produce pre-trained language identifiers targeted specifically at Twitter messages. `LDIG` is a Twitter-specific LI tool with built-in models for 19 languages.[19] It uses a document representation based on tries (Okanohara and Tsujii, 2009). The algorithm is a LR classifier using all possible substrings of the data, which is important to maximize the available information from the relatively short Twitter messages.

Lui and Baldwin (2014) provides a comparison of 8 off-the-shelf language identifiers applied without re-training to Twitter messages. One issue they report is that comparing the accuracy of off-the-shelf systems is difficult because of the different subset of languages supported by each system, which may also not fully cover the languages present in the target data. The authors choose to compare accuracy over the full set of languages, arguing that this best reflects the likely use-case of applying an off-the-shelf LI system to new data. They find that the best individual systems are `ChromeCLD`, `langid.py` and `LangDetect`, but that slightly higher accuracy can be attained by a simple voting-based ensemble classifier involving these three systems.

In addition to this, commercial or other closed-source language identifiers and language identifier services exist, of which we name a few. The Polyglot 3000[20] and Lextek Language Identifier[21] are standalone language identifiers for Windows. Open Xerox Language Identifier[22] is a web service with available REST and SOAP APIs.

13. `https://github.com/carrotsearch/langid-java` (last updated on June 2013)
14. `https://github.com/saffsd/langid.c` (last updated on September 2017)
15. `https://github.com/saffsd/langid.js` (last updated on July 2014)
16. `https://sourceforge.net/projects/la-strings/` (last updated on February 2018)
17. `https://github.com/indix/whatthelang` (last updated on November 2017)
18. `https://github.com/martin-majlis/YALI` (last updated on May 2014)
19. `https://github.com/shuyo/ldig` (last updated on July 2013)
20. `http://www.polyglot3000.com`
21. `http://www.lextek.com/langid/li/`
22. `https://open.xerox.com/Services/LanguageIdentifier`

10. Research Directions and Open Issues in LI

Several papers have catalogued open issues in LI (Sibun and Reynar, 1996; Xia et al., 2010; Hughes et al., 2006; da Silva and Lopes, 2006a; Baldwin and Lui, 2010b; Botha and Barnard, 2012; Malmasi et al., 2016). Some of the issues, such as text representation (Section 5) and choice of algorithm (Section 6), have already been covered in detail in this survey. In this section, we synthesize the remaining issues into a single section, and also add new issues that have not been discussed in previous work. For each issue, we review related work and suggest promising directions for future work.

10.1 Text Preprocessing

Text preprocessing (also known as normalization) is an umbrella term for techniques where an automatic transformation is applied to text before it is presented to a classifier. The aim of such a process is to eliminate sources of variation that are expected to be confounding factors with respect to the target task. Text preprocessing is slightly different from data cleaning, as data cleaning is a transformation applied only to training data, whereas normalization is applied to both training and test data. Hughes et al. (2006) raise text preprocessing as an outstanding issue in LI, arguing that its effects on the task have not been sufficiently investigated. In this section, we summarize the normalization strategies that have been proposed in the LI literature.

Case folding is the elimination of capitalization, replacing characters in a text with either their lower-case or upper-case forms. Basic approaches generally map between `[a-z]` and `[A-Z]` in the ASCII encoding, but this approach is insufficient for extended Latin encodings, where diacritics must also be appropriately handled. A resource that makes this possible is the Unicode Character Database (UCD)[23] which defines uppercase, lowercase and titlecase properties for each character, enabling automatic case folding for documents in a Unicode encoding such as UTF-8.

Range compression is the grouping of a range of characters into a single logical set for counting purposes, and is a technique that is commonly used to deal with the sparsity that results from character sets for ideographic languages, such as Chinese, that may have thousands of unique "characters", each of which is observed with relatively low frequency. Simões et al. (2014) use such a technique where all characters in a given range are mapped into a single "bucket", and the frequency of items in each bucket is used as a feature to represent the document. Byte-level representations of encodings that use multi-byte sequences to represent codepoints achieve a similar effect by "splitting" codepoints. In encodings such as UTF-8, the codepoints used by a single language are usually grouped together in "code planes", where each codepoint in a given code plane shares the same upper byte. Thus, even though the distribution over codepoints may be quite sparse, when the byte-level representation uses byte sequences that are shorter than the multi-byte sequence of a codepoint, the shared upper byte will be predictive of specific languages.

Cleaning may also be applied, where heuristic rules are used to remove some data that is perceived to hinder the accuracy of the language identifier. For example, Suzuki et al. (2002) identify HTML entities as a candidate for removal in document cleaning, on the basis

23. `http://www.unicode.org/ucd/`

that classifiers trained on data which does not include such entities may drop in accuracy when applied to raw HTML documents. `ChromeCLD` includes heuristics such as expanding HTML entities, deleting digits and punctuation, and removing SGML-like tags. Similarly, `LangDetect` also removes "language-independent characters" such as numbers, symbols, URLs, and email addresses. It also removes words that are all-capitals and tries to remove other acronyms and proper names using heuristics.

In the domain of Twitter messages, Tromp and Pechenizkiy (2011) remove links, usernames, smilies, and hashtags (a Twitter-specific "tagging" feature), arguing that these entities are language independent and thus should not feature in the model. Xafopoulos et al. (2004) address LI of web pages, and report removing HTML formatting, and applying stopping using a small stopword list. Takçı and Ekinci (2012) carry out LI experiments on the ECI multilingual corpus and report removing punctuation, space characters, and digits.

The idea of preprocessing text to eliminate domain-specific "noise" is closely related to the idea of learning domain-independent characteristics of a language (Lui and Baldwin, 2011). One difference is that normalization is normally heuristic-driven, where a manually-specified set of rules is used to eliminate unwanted elements of the text, whereas domain-independent text representations are data-driven, where text from different sources is used to identify the characteristics that a language shares between different sources. Both approaches share conceptual similarities with problems such as content extraction for web pages. In essence, the aim is to isolate the components of the text that actually represent language, and suppress the components that carry other information. One application is the language-aware extraction of text strings embedded in binary files, which has been shown to perform better than conventional heuristic approaches (Brown, 2012). Future work in this area could focus specifically on the application of language-aware techniques to content extraction, using models of language to segment documents into textual and non-textual components. Such methods could also be used to iteratively improve LI itself by improving the quality of training data.

10.2 Orthography and Transliteration

LI is further complicated when we consider that some languages can be written in different orthographies (e.g. Bosnian and Serbian can be written in both Latin and Cyrillic script). Transliteration is another phenomenon that has a similar effect, whereby phonetic transcriptions in another script are produced for particular languages. These transcriptions can either be standardized and officially sanctioned, such as the use of *Hanyu Pinyin* for Chinese, or may also emerge irregularly and organically as in the case of *arabizi* for Arabic (Yaghan, 2008). Hughes et al. (2006) identify variation in the encodings and scripts used by a given language as an open issue in LI, pointing out that early work tended to focus on languages written using a romanized script, and suggesting that dealing with issues of encoding and orthography adds substantial complexity to the task. Suzuki et al. (2002) discuss the relative difficulties of discriminating between languages that vary in any combination of encoding, script and language family, and give examples of pairs of languages that fall into each category.

LI across orthographies and transliteration is an area that has not received much attention in work to date, but presents unique and interesting challenges that are suitable

targets for future research. An interesting and unexplored question is whether it is possible to detect that documents in different encodings or scripts are written in the same language, or what language a text is transliterated from, without any a-priori knowledge of the encoding or scripts used. One possible approach to this could be to take advantage of standard orderings of alphabets in a language – the pattern of differences between adjacent characters should be consistent across encodings, though whether this is characteristic of any given language requires exploration.

10.3 Supporting Low-Resource Languages

Hughes et al. (2006) paint a fairly bleak picture of the support for low-resource languages in automatic LI. This is supported by the arguments of Xia et al. (2010) who detail specific issues in building hugely multilingual datasets. Abney and Bird (2010) also specifically called for research into automatic LI for low-density languages. Ethnologue (Simons and Fennig, 2017) lists a total of 7099 languages. Xia et al. (2010) describe the Ethnologue in more detail, and discuss the role that LI plays in other aspects of supporting minority languages, including detecting and cataloging resources. The problem is circular: LI methods are typically supervised, and need training data for each language to be covered, but the most efficient way to recover such data is through LI methods.

A number of projects are ongoing with the specific aim of gathering linguistic data from the web, targeting as broad a set of languages as possible. One such project is the aforementioned ODIN (Xia et al., 2009; Lewis and Xia, 2010), which aims to collect parallel snippets of text from Linguistics articles published on the web. ODIN specifically targets articles containing Interlinear Glossed Text (IGT), a semi-structured format for presenting text and a corresponding gloss that is commonly used in Linguistics.

Other projects that exist with the aim of creating text corpora for under-resourced languages by crawling the web are the Crúbadán project (Scannell, 2007) and SeedLing (Emerson et al., 2014). The Crúbadán crawler uses seed data in a target language to generate word lists that in turn are used as queries for a search engine. The returned documents are then compared with the seed resource via an automatic language identifier, which is used to eliminate false positives. Scannell (2007) reports that corpora for over 400 languages have been built using this method. The SeedLing project crawls texts from several web sources which has resulted in a total of 1451 languages from 105 language families. According to the authors, this represents 19% of the world's languages.

Much recent work on multilingual documents (Section 10.6) has been done with support for minority languages as a key goal. One of the common problems with gathering linguistic data from the web is that the data in the target language is often embedded in a document containing data in another language. This has spurred recent developments in text segmentation by language and word-level LI. Lui et al. (2014a) present a method to detect documents that contain text in more than one language and identify the languages present with their relative proportions in the document. The method is evaluated on real-world data from a web crawl targeted to collect documents for specific low-density languages.

LI for low-resource languages is a promising area for future work. One of the key questions that has not been clearly answered is how much data is needed to accurately model a language for purposes of LI. Work to date suggests that there may not be a simple

answer to this question as accuracy varies according to the number and variety of languages modeled (Baldwin and Lui, 2010b), as well as the diversity of data available to model a specific language (Lui and Baldwin, 2011).

10.4 Number of Languages

Early research in LI tended to focus on a very limited number of languages (sometimes as few as 2). This situation has improved somewhat with many current off-the-shelf language identifiers supporting on the order of 50–100 languages (Section 9). The standout in this regard is Brown (2014), supporting 1311 languages in its default configuration. However, evaluation of the identifier of Brown (2013) on a different domain found that the system suffered in terms of accuracy because it detected many languages that were not present in the test data (Lui and Baldwin, 2014).

Lewis and Xia (2010) describe the construction of web crawlers specifically targeting IGT, as well as the identification of the languages represented in the IGT snippets. LI for thousands of languages from very small quantities of text is one of the issues that they have had to tackle. They list four specific challenges for LI in ODIN: (1) the large number of languages; (2) "unseen" languages that appear in the test data but not in training data; (3) short target sentences; and (4) (sometimes inconsistent) transliteration into Latin text. Their solution to this task is to take advantage of a domain-specific feature: they assume that the name of the language that they are extracting must appear in the document containing the IGT, and hence treat this as a co-reference resolution problem. They report that this approach significantly outperforms the text-based LI approach in this particular problem setting.

An interesting area to explore is the trade-off between the number of languages supported and the accuracy per-language. From existing results it is not clear if it is possible to continue increasing the number of languages supported without adversely affecting the average accuracy, but it would be useful to quantify if this is actually the case across a broad range of text sources. Table 13 lists the articles where the LI with more than 30 languages has been investigated.

10.5 "Unseen" Languages and Unsupervised LI

"Unseen" languages are languages that we do not have training data for but may nonetheless be encountered by a LI system when applied to real-world data. Dealing with languages for which we do not have training data has been identified as an issue by Hughes et al. (2006) and has also been mentioned by Xia et al. (2009) as a specific challenge in harvesting linguistic data from the web. Elfardy and Diab (2012) use an unlabeled training set with a labeled evaluation set for token-level code switching identification between Modern Standard Arabic (MSA) and dialectal Arabic. They utilize existing dictionaries and also a morphological analyzer for MSA, so the system is supported by extensive external knowledge sources. The possibility to use unannotated training material is nonetheless a very useful feature.

Some authors have attempted to tackle the unseen language problem through attempts at unsupervised labeling of text by language. Mather (1998) uses an unsupervised clustering algorithm to separate a multilingual corpus into groups corresponding to languages.

Reference	# Lang	Reference	# Lang
Brown (2014)	1311	Brown (2013)	1100
Brown (2012)	923	Xia et al. (2009)	c. 600
Rodrigues (2012)	372	King and Dehdari (2008)	300
Jauhiainen et al. (2015c)	285	Jauhiainen et al. (2017a)	285
Vatanen et al. (2010)	281	Yamaguchi and Tanaka-Ishii (2012)	200+
Cazamias et al. (2015)	200	Chew et al. (2011)	182
Lui (2014)	143	Kocmi and Bojar (2017)	136
Majliš (2011)	122	Jauhiainen (2010)	103
Majliš (2012)	90	Lui and Baldwin (2011)	89
Baldwin and Lui (2010a)	74	Chew et al. (2009)	68
Baldwin and Lui (2010b)	67	Lui and Baldwin (2012)	67
Lui and Baldwin (2014)	65	Goldszmidt et al. (2013)	52
Chen and Maison (2003)	48	Lui et al. (2014a)	44
Singh (2006)	39	Cowie et al. (1999)	34
Ludovik and Zacharski (1999)	34	Hammarström (2007)	32
Abainia et al. (2014)	32	King and Abney (2013)	31

Table 13: Empirical evaluations with more than 30 languages.

She uses singular value decomposition (SVD) to first identify the words that discriminate between documents and then to separate the terms into highly correlating groups. The documents grouped together by these discriminating terms are merged and the process is repeated until the wanted number of groups (corresponding to languages) is reached. Biemann and Teresniak (2005) also presents an approach to unseen language problem, building graphs of co-occurrences of words in sentences, and then partitioning the graph using a custom graph-clustering algorithm which labels each word in the cluster with a single label. The number of labels is initialized to be the same as the number of words, and decreases as the algorithm is recursively applied. After a small number of iterations (the authors report 20), the labels become relatively stable and can be interpreted as cluster labels. Smaller clusters are then discarded, and the remaining clusters are interpreted as groups of words for each language. Shiells and Pham (2010) compared the Chinese Whispers algorithm of Biemann and Teresniak (2005) and Graclus clustering on unsupervised Tweet LI. They conclude that Chinese Whispers is better suited to LI. Selamat and Ng (2008) used Fuzzy ART NNs for unsupervised language clustering for documents in Arabic, Persian, and Urdu. In Fuzzy ART, the clusters are also dynamically updated during the identification process.

Amine et al. (2010) also tackle the unseen language problem through clustering. They use a character n-gram representation for text, and a clustering algorithm that consists of an initial k-means phase, followed by particle-swarm optimization. This produces a large number of small clusters, which are then labeled by language through a separate step. Wan (2016) used co-occurrences of words with k-means clustering in word-level unsupervised LI. They used a Dirichlet process Gaussian mixture model ("DPGMM"), a non-parametric variant of a GMM, to automatically determine the number of clusters, and manually labeled the language of each cluster. Poulston et al. (2017) also used k-means clustering, and Alfter (2015) used the x-means clustering algorithm in a custom framework. Lin et al.

(2014) utilized unlabeled data to improve their LI system by using a CRF autoencoder, unsupervised word embeddings, and word lists.

A different partial solution to the issue of unseen languages is to design the classifier to be able to output "unknown" as a prediction for language. This helps to alleviate one of the problems commonly associated with the presence of unseen languages – classifiers without an "unknown" facility are forced to pick a language for each document, and in the case of unseen languages, the choice may be arbitrary and unpredictable (Biemann and Teresniak, 2005). When LI is used for filtering purposes, i.e. to select documents in a single language, this mislabeling can introduce substantial noise into the data extracted; furthermore, it does not matter what or how many unseen languages there are, as long as they are consistently rejected. Therefore the "unknown" output provides an adequate solution to the unseen language problem for purposes of filtering.

The easiest way to implement unknown language detection is through thresholding. Most systems internally compute a score for each language for an unknown text, so thresholding can be applied either with a global threshold (Cowie et al., 1999), a per-language threshold (Suzuki et al., 2002), or by comparing the score for the top-scoring N-languages. The problem of unseen languages and open-set recognition was also considered by Malmasi and Dras (2015b), Zampieri et al. (2015a), and Malmasi (2017). Malmasi (2017) experiments with one-class classification ("OCC") and reaches an F-score on 98.9 using OC-SVMs (SVMs trained only with data from one language) to discriminate between 10 languages.

Another possible method for unknown language detection that has not been explored extensively in the literature, is the use of non-parametric mixture models based on Hierarchical Dirichlet Processes ("HDP"). Such models have been successful in topic modeling, where an outstanding issue with the popular LDA model is the need to specify the number of topics in advance. Lui et al. (2014a) introduced an approach to detecting multilingual documents that uses a model very similar to LDA, where languages are analogous to topics in the LDA model. Using a similar analogy, an HDP-based model may be able to detect documents that are written in a language that is not currently modeled by the system. Voss et al. (2014) used LDA to cluster unannotated tweets. Recently Zhang et al. (2016) used LDA in unsupervised sentence-level LI. They manually identified the languages of the topics created with LDA. If there were more topics than languages then the topics in the same language were merged.

Filtering, a task that we mentioned earlier in this section, is a very common application of LI, and it is therefore surprising that there is little research on filtering for specific languages. Filtering is a limit case of LI with unseen languages, where all languages but one can be considered unknown. Future work could examine how useful different types of negative evidence are for filtering – if we want to detect English documents, e.g., are there empirical advantages in having distinct models of Italian and German (even if we don't care about the distinction between the two languages), or can we group them all together in a single "negative" class? Are we better off including as many languages as possible in the negative class, or can we safely exclude some?

10.6 Multilingual Documents

Multilingual documents are documents that contain text in more than one language. In constructing the hrWac corpus, Stupar et al. (2011) found that 4% of the documents they collected contained text in more than one language. Martins and Silva (2005) report that web pages in many languages contain formulaic strings in English that do not actually contribute to the content of the page, but may nonetheless confound attempts to identify multilingual documents. Recent research has investigated how to make use of multilingual documents from sources such as web crawls (King and Abney, 2013), forum posts (Nguyen and Dogruöz, 2013), and microblog messages (Ling et al., 2013). However, most LI methods assume that a document contains text from a single language, and so are not directly applicable to multilingual documents.

Handling of multilingual documents has been named as an open research question (Hughes et al., 2006). Most NLP techniques presuppose monolingual input data, so inclusion of data in foreign languages introduces noise, and can degrade the performance of NLP systems. Automatic detection of multilingual documents can be used as a prefiltering step to improve the quality of input data. Detecting multilingual documents is also important for acquiring linguistic data from the web, and has applications in mining bilingual texts for statistical MT from online resources (Ling et al., 2013), or to study code-switching phenomena in online communications. There has also been interest in extracting text resources for low-density languages from multilingual web pages containing both the low-density language and another language such as English.

The need to handle multilingual documents has prompted researchers to revisit the granularity of LI. Many researchers consider document-level LI to be relatively easy, and that sentence-level and word-level LI are more suitable targets for further research. However, word-level and sentence-level tokenization are not language-independent tasks, and for some languages are substantially harder than others (Peng et al., 2004).

`Linguini` (Prager, 1999) is a language identifier that supports identification of multilingual documents. The system is based on a vector space model using cosine similarity. LI for multilingual documents is performed through the use of *virtual mixed languages*. Prager (1999) shows how to construct vectors representative of particular combinations of languages independent of the relative proportions, and proposes a method for choosing combinations of languages to consider for any given document. One weakness of this approach is that for exhaustive coverage, this method is factorial in the number of languages, and as such intractable for a large set of languages. Furthermore, calculating the parameters for the virtual mixed languages becomes infeasibly complex for mixtures of more than 3 languages.

As mentioned previously, Lui et al. (2014a) propose an LDA-inspired LI method for multilingual documents that is able to identify that a document is multilingual, identify the languages present and estimate the relative proportions of the document written in each language. To remove the need to specify the number of topics (or in this case, languages) in advance, Lui et al. (2014a) use a greedy heuristic that attempts to find the subset of languages that maximizes the posterior probability of a target document. One advantage of this approach is that it is not constrained to 3-language combinations like the method of

Prager (1999). Language set identification has also been considered by Suzuki et al. (2002), Jauhiainen et al. (2015c), and Pla and Hurtado (2015, 2017).

To encourage further research on LI for multilingual documents, in the aforementioned shared task hosted by the Australiasian Language Technology Workshop 2010, discussed in Section 7.3, participants were required to predict the language(s) present in a held-out test set containing monolingual and bilingual documents (Baldwin and Lui, 2010a). The dataset was prepared using data from Wikipedia, and bilingual documents were produced using a segment from an article in one language and a segment from the equivalent article in another language. Equivalence between articles was determined using the cross-language links embedded within each Wikipedia article.[24] The winning entry (Tran et al., 2010) first built monolingual models from multilingual training data, and then applied them to a chunked version of the test data, making the final prediction a function of the prediction over chunks.

Another approach to handling multilingual documents is to attempt to segment them into contiguous monolingual segments. In addition to identifying the languages present, this requires identifying the locations of boundaries in the text which mark the transition from one language to another. Several methods for supervised language segmentation have been proposed. Cowie et al. (1999) generalized a LI algorithm for monolingual documents by adding a dynamic programming algorithm based on a simple Markov model of multilingual documents. More recently, multilingual LI algorithms have also been presented by Jhamtani et al. (2014), Minocha and Tyers (2014), Pethö and Mózes (2014), Ullman (2014), and King et al. (2015).

10.7 Short Texts

LI of short strings is known to be challenging for existing LI techniques. Mandl et al. (2006) tested four different classification methods, and found that all have substantially lower accuracy when applied to texts of 25 characters compared with texts of 125 characters. These findings were later strengthened, for example, by Vatanen et al. (2010) and Jauhiainen et al. (2017a).

Hammarström (2007) describes a method specifically targeted at short texts that augments a dictionary with an affix table, which was tested over synthetic data derived from a parallel bible corpus. Vatanen et al. (2010) focus on messages of 5–21 characters, using n-gram language models over data drawn the from Universal Declaration of Human Rights (UDHR). We would expect that generic methods for LI of short texts should be effective in any domain where short texts are found, such as search engine queries or microblog messages. However, Hammarström (2007) and Vatanen et al. (2010) both only test their systems in a single domain: bible texts in the former case, and texts from the UDHR in the latter case. Other research has shown that LI results do not trivially generalize across domains (Baldwin and Lui, 2010b), and found that LI in UDHR documents is relatively easy (Yamaguchi and Tanaka-Ishii, 2012). For both bible and UDHR data, we expect that the linguistic content is relatively grammatical and well-formed, an expectation that does not carry across to domains such as search engine queries and microblogs. Another "short

24. Note that such articles are not necessarily direct translations but rather articles about the same topic written in different languages.

text" domain where LI has been studied is LI of proper names. Häkkinen and Tian (2001) identify this as an issue. Konstantopoulos (2007) found that LI of names is more accurate than LI of generic words of equivalent length.

Bergsma et al. (2012) raise an important criticism of LI work on Twitter messages to date: only a small number of European languages has been considered. Bergsma et al. (2012) expand the scope of LI for Twitter, covering nine languages across Cyrillic, Arabic and Devanagari scripts. Lui and Baldwin (2014) expand the evaluation further, introducing a dataset of language-labeled Twitter messages across 65 languages constructed using a semi-automatic method that leverages user identity to avoid inducing a bias in the evaluation set towards messages that existing systems are able to identify correctly. Lui and Baldwin (2014) also test a 1300-language model based on Brown (2013), but find that it performs relatively poorly in the target domain due to a tendency to over-predict low-resource languages.

Work has also been done on LI of single words in a document, where the task is to label each word in the document with a specific language. Work to date in this area has assumed that word tokenization can be carried out on the basis of whitespace. Singh and Gorla (2007) explore word-level LI in the context of segmenting a multilingual document into monolingual segments. Other work has assumed that the languages present in the document are known in advance.

Conditional random fields ("CRFs": Lafferty et al. (2001)) are a sequence labeling method most often used in LI for labeling the language of individual words in a multilingual text. CRFs can be thought of as a finite state model with probabilistic transition probabilities optimised over pre-defined cliques. They can use any observations made from the test document as features, including language labels given by monolingual language identifiers for words. King and Abney (2013) used a CRF trained with generalized expectation criteria, and found it to be the most accurate of all methods tested (NB, LR, HMM, CRF) at word-level LI. King and Abney (2013) introduce a technique to estimate the parameters using only monolingual data, an important consideration as there is no readily-available collection of manually-labeled multilingual documents with word-level annotations. Nguyen and Dogruöz (2013) present a two-pass approach to processing Turkish-Dutch bilingual documents, where the first pass labels each word independently and the second pass uses the local context of a word to further refine the predictions. Nguyen and Dogruöz (2013) achieved 97,6% accuracy on distinguishing between the two languages using a linear-chain CRF. Clematide and Makarov (2017) are the only ones so far to use a CRF for LI of monolingual texts. With a CRF, they attained a higher F-score in German dialect identification than NB or an ensemble consisting of NB, CRF, and SVM. Lately CRFs were also used for LI by Dongen (2017) and Samih (2017). Giwa and Davel (2013) investigate LI of individual words in the context of code switching. They find that smoothing of n-gram models substantially improves accuracy of a language identifier based on a NB classifier when applied to individual words.

10.8 Similar Languages, Language Varieties, and Dialects

While one line of research into LI has focused on pushing the boundaries of how many languages are supported simultaneously by a single system (Xia et al., 2010; Brown, 2012,

2013), another has taken a complementary path and focused on LI in groups of similar languages. Research in this area typically does not make a distinction between languages, varieties and dialects, because such terminological differences tend to be politically rather than linguistically motivated (Clyne, 1992; Xia et al., 2010; Zampieri and Gebre, 2012), and from an NLP perspective the challenges faced are very similar.

LI for closely-related languages, language varieties, and dialects has been studied for Malay–Indonesian (Ranaivo-Malançon, 2006), Indian languages (Murthy and Kumar, 2006), South Slavic languages (Ljubešić et al., 2007; Tiedemann and Ljubešić, 2012; Ljubešić and Kranjcić, 2014, 2015), Serbo-Croatian dialects (Zecevic and Vujicic-Stankovic, 2013), English varieties (Lui and Cook, 2013; Simaki et al., 2017), Dutch–Flemish (van der Lee and Bosch, 2017), Dutch dialects (including a temporal dimension) (Trieschnigg et al., 2012), German Dialects (Hollenstein and Aepli, 2015) Mainland–Singaporean–Taiwanese Chinese (Huang and Lee, 2008), Portuguese varieties (Zampieri and Gebre, 2012; Zampieri et al., 2016), Spanish varieties (Zampieri et al., 2013; Maier and Gómez-Rodríguez, 2014), French varieties (Mokhov, 2010a,b; Diwersy et al., 2014), languages of the Iberian Peninsula (Zubiaga et al., 2014), Romanian dialects (Ciobanu and Dinu, 2016), and Arabic dialects (Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2014; Tillmann et al., 2014; Sadat et al., 2014b; Wray, 2018), the last of which we discuss in more detail in this section. As to off-the-shelf tools which can identify closely-related languages, Zampieri and Gebre (2014) released a LI system trained to identify 27 languages, including 10 language varieties. Closely-related languages, language varieties, and dialects have also been the focus of a number of shared tasks in recent years as discussed in Section 7.3.

Similar languages are a known problem for existing language identifiers (Ranaivo-Malançon, 2006; Zampieri, 2013). Suzuki et al. (2002) identify language pairs from the same language family that also share a common script and the same encoding, as the most difficult to discriminate. Tiedemann and Ljubešić (2012) report that `TextCat` achieves only 45% accuracy when trained and tested on 3-way Bosnian/Serbian/Croatian dataset. Lui and Cook (2013) found that LI methods are not competitive with conventional word-based document categorization methods in distinguishing between national varieties of English. Ranaivo-Malançon (2006) reports that a character trigram model is able to distinguish Malay/Indonesian from English, French, German, and Dutch, but handcrafted rules are needed to distinguish between Malay and Indonesian. One kind of rule is the use of "exclusive words" that are known to occur in only one of the languages. A similar idea is used by Tiedemann and Ljubešić (2012), in automatically learning a "blacklist" of words that have a strong negative correlation with a language – i.e. their presence implies that the text is *not* written in a particular language. In doing so, they achieve an overall accuracy of 98%, far surpassing the 45% of `TextCat`. Brown (2013) also adopts such "discriminative training" to make use of negative evidence in LI.

Zampieri (2013) observed that general-purpose approaches to LI typically use a character n-gram representation of text, but successful approaches for closely-related languages, varieties, and dialects seem to favor a word-based representation or higher-order n-grams (e.g. 4-grams, 5-grams, and even 6-grams) that often cover whole words (Huang and Lee, 2008; Tiedemann and Ljubešić, 2012; Lui and Cook, 2013; Goutte et al., 2016). The study compared character n-grams with word-based representations for LI over varieties of Spanish,

Portuguese and French, and found that word-level models performed better for varieties of Spanish, but character n-gram models perform better in the case of Portuguese and French.

To train accurate and robust LI systems that discriminate between language varieties or similar languages, models should ideally be able to capture not only lexical but more abstract systemic differences between languages. One way to achieve this, is by using features that use de-lexicalized text representations (e.g. by substituting named entities or content words by placeholders), or at a higher level of abstraction, using POS tags or other morphosyntactic information (Zampieri et al., 2013; Lui et al., 2014b; Bestgen, 2017), or even adversarial machine learning to modify the learned representations to remove such artefacts (Li et al., 2018). Finally, an interesting research direction could be to combine work on closely-related languages with the analysis of regional or dialectal differences in language use (Peirsman et al., 2010; Anstein, 2013; Doyle, 2014; Diwersy et al., 2014).

In recent years, there has been a significant increase of interest in the computational processing of Arabic. This is evidenced by a number of research papers in several NLP tasks and applications including the identification/discrimination of Arabic dialects (Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2014). Arabic is particularly interesting for researchers interested in language variation due to the fact that the language is often in a diaglossic situation, in which the standard form (Modern Standard Arabic or "MSA") coexists with several regional dialects which are used in everyday communication.

Among the studies published on the topic of Arabic LI, Elfardy and Diab (2013) proposed a supervised approach to distinguish between MSA and Egyptian Arabic at the sentence level, and achieved up to 85.5% accuracy over an Arabic online commentary dataset (Zaidan and Callison-Burch, 2011). Tillmann et al. (2014) achieved higher results over the same dataset using a linear-kernel SVM classifier.

Zaidan and Callison-Burch (2014) compiled a dataset containing MSA, Egyptian Arabic, Gulf Arabic and Levantine Arabic, and used it to investigate three classification tasks: (1) MSA and dialectal Arabic; (2) four-way classification – MSA, Egyptian Arabic, Gulf Arabic, and Levantine Arabic; and (3) three-way classification – Egyptian Arabic, Gulf Arabic, and Levantine Arabic.

Salloum et al. (2014) explores the use of sentence-level Arabic dialect identification as a pre-processor for MT, in customizing the selection of the MT model used to translate a given sentence to the dialect it uses. In performing dialect-specific MT, the authors achieve an improvement of 1.0% BLEU score compared with a baseline system which does not differentiate between Arabic dialects.

Finally, in addition to the above-mentioned dataset of Zaidan and Callison-Burch (2011), there are a number of notable multi-dialect corpora of Arabic: a multi-dialect corpus of broadcast speeches used in the ADI shared task (Ali et al., 2016); a multi-dialect corpus of (informal) written Arabic containing newspaper comments and Twitter data (Cotterell and Callison-Burch, 2014); a parallel corpus of 2,000 sentences in MSA, Egyptian Arabic, Tunisian Arabic, Jordanian Arabic, Palestinian Arabic, and Syrian Arabic, in addition to English (Bouamor et al., 2014); a corpus of sentences in 18 Arabic dialects (corresponding to 18 different Arabic-speaking countries) based on data manually sourced from web forums (Sadat et al., 2014b); and finally two recently compiled multi-dialect corpora containing microblog posts from Twitter (Elgabou and Kazakov, 2017; Alshutayri and Atwell, 2017).

While not specifically targeted at identifying language varieties, Jurgens et al. (2017) made the critical observation that when naively trained, LI systems tend to perform most poorly over language varieties from the lowest socio-economic demographics (focusing particularly on the case of English), as they tend to be most under-represented in training corpora. If, as a research community, we are interested in the social equitability of our systems, it is critical that we develop datasets that are truly representative of the global population, to better quantify and remove this effect. To this end, Jurgens et al. (2017) detail a method for constructing a more representative dataset, and demonstrate the impact of training on such a dataset in terms of alleviating socio-economic bias.

10.9 Domain-specific LI

One approach to LI is to build a generic language identifier that aims to correctly identify the language of a text without any information about the source of the text. Some work has specifically targeted LI across multiple domains, learning characteristics of languages that are consistent between different sources of text (Lui and Baldwin, 2011). However, there are often domain-specific features that are useful for identifying the language of a text. In this survey, our primary focus has been on LI of digitally-encoded text, using only the text itself as evidence on which to base the prediction of the language. Within a text, there can sometimes be domain-specific peculiarities that can be used for LI. For example, Mayer (2012) investigates LI of user-to-user messages in the eBay e-commerce portal. He finds that using only the first two and last two words of a message is sufficient for identifying the language of a message.

11. Conclusions

This article has presented a comprehensive survey on language identification of digitally-encoded text. We have shown that LI is a rich, complex, and multi-faceted problem that has engaged a wide variety of research communities. LI accuracy is critical as it is often the first step in longer text processing pipelines, so errors made in LI will propagate and degrade the performance of later stages. Under controlled conditions, such as limiting the number of languages to a small set of Western European languages and using long, grammatical, and structured text such as government documents as training data, it is possible to achieve near-perfect accuracy. This led many researchers to consider LI a solved problem, as argued by McNamee (2005). However, LI becomes much harder when taking into account the peculiarities of real-world data, such as very short documents (e.g. search engine queries), non-linguistic "noise" (e.g. HTML markup), non-standard use of language (e.g. as seen in social media data), and mixed-language documents (e.g. forum posts in multilingual web forums).

Modern approaches to LI are generally data-driven and are based on comparing new documents with models of each target language learned from data. The types of models as well as the sources of training data used in the literature are diverse, and work to date has not compared and evaluated these in a systematic manner, making it difficult to draw broader conclusions about what the "best" method for LI actually is. We have attempted to synthesize results to date to identify a set of LI "best practices", but these should be

treated as guidelines and should always be considered in the broader context of a target application.

Existing work on LI serves to illustrate that the scope and depth of the problem are much greater than they may first appear. In Section 10, we discussed open issues in LI, identifying the key challenges, and outlining opportunities for future research. Far from being a solved problem, aspects of LI make it an archetypal learning task with subtleties that could be tackled by future work on supervised learning, representation learning, multi-task learning, domain adaptation, multi-label classification and other subfields of machine learning. We hope that this paper can serve as a reference point for future work in the area, both for providing insight into work to date, as well as pointing towards the key aspects that merit further investigation.

Acknowledgments

This research was supported in part by the Australian Research Council, the Kone Foundation and the Academy of Finland. We would like to thank Kimmo Koskenniemi for many valuable discussions and comments concerning the early phases of the features and the methods sections.

References

Kheireddine Abainia, Siham Ouamour, and Halim Sayoud. Robust Language Identification of Noisy Texts - Proposal of Hybrid Approaches. In *25th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 228–232, Munich, Germany, 2014.

Kheireddine Abainia, Siham Ouamour, and Halim Sayoud. Effective Language Identification of Forum Texts Based on Statistical Approaches. *Information Processing and Management*, 52:491–512, 2016.

Steven Abney and Steven Bird. The Human Language Project: Building a Universal Corpus of the World's Languages. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 88–97, Los Angeles, USA, 2010.

Judit Ács, László Grad-Gyenge, Thiago Bruno, and Rodriguez de Rezende Oliveira. A Two-level Classifier for Discriminating similar Languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 73–77, Hissar, Bulgaria, 2015.

Gary Adams and Philip Resnik. A Language Identification Application Built on the Java Client/server Platform. In *Proceedings of the ACL/EACL'97 Workshop on From Research to Commercial Applications: Making NLP Work in Practice*, pages 43–47, Madrid, Spain, 1997.

Wafia Adouane. Automatic Detection of Underresourced Languages: Dialectal Arabic Short Texts. Master's thesis, University of Gothenburg, Gothenburg, Sweden, 2016.

Wafia Adouane and Simon Dobnik. Identification of Languages in Algerian Arabic Multilingual Documents. In *Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP 2017)*, pages 1–8, Valencia, Spain, 2017.

Wafia Adouane, Nasredine Semmar, and Richard Johansson. Romanized Berber and Romanized Arabic Automatic Language Identification Using Machine Learning. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 53–61, Osaka, Japan, 2016a.

Wafia Adouane, Nasredine Semmar, and Richard Johansson. Arabicized and Romanized Berber Automatic Identification. In *Proceedings of the International Conference on Information and Communication Technologies for Amazingh (TICAM 2016)*, Rabat, Morocco, 2016b. IRCAM.

Wafia Adouane, Nasredine Semmar, and Richard Johansson. ASIREM Participation at the Discriminating Similar Languages Shared Task 2016. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 163–169, Osaka, Japan, 2016c.

Wafia Adouane, Nasredine Semmar, Richard Johansson, and Victoria Bobicev. Automatic Detection of Arabicized Berber and Arabic Varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 63–72, Osaka, Japan, 2016d.

Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert. Language Identification from Text Using N-gram Based Cumulative Frequency Addition. In *Proceedings of Student/Faculty Research Day*, pages 12.1–12.8, CSIS, Pace University, New York, USA, 2004.

Malepati Bala Siva Sai Akhil and J. Abhishek. Language Identification, Transliteration and Resolving Common Words Ambiguity in a Pair of Languages: Shared Task on Transliterated Search. In *Working Notes of Shared Task on Transliterated Search at Forum for Information Retrieval Evaluation (FIRE'14)*, Bangalore, India, 2014.

Liliya Akhtyamova, John Cardiff, and Andrey Ignatov. Twitter Author Profiling Using Word Embeddings and Logistic Regression - Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Nicholas Akosu and Ali Selamat. A Dynamic Model Selection Algorithm for Language Identification of Under-resourced Languages. *International Journal of Digital Content Technology and its Applications (JDCTA)*, 8, 2014.

Mohamed Al-Badrashiny and Mona T. Diab. LILI: A Simple Language Independent Approach for Language Identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers*, pages 1211–1219, Osaka, Japan, 2016.

Mohamed Al-Badrashiny, Heba Elfardy, and Mona Diab. AIDA2: A Hybrid Approach for Token and Sentence Level Dialect Identification in Arabic. In *Proceedings of the 19th Conference on Computational Language Learning*, pages 42–51, Beijing, China, 2015.

Beatrice Alex. An Unsupervised System for Identifying English Inclusions in German Text. In *Proceedings of the Student Research Workshop, ACL-05*, pages 133–138, Ann Arbor, Michigan, USA, 2005.

Beatrice Alex, Amit Dubey, and Frank Keller. Using Foreign Inclusion Detection to Improve Parsing Performance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007 (EMNLP-CoNLL 2007)*, pages 151–160, Prague, Czech Republic, 2007.

David Alfter. Language Segmentation. Master's thesis, Universität Trier, Trier, Germany, 2015.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. Automatic Dialect Detection in Arabic Broadcast Speech. In *Proceedings of Interspeech 2016*, pages 2934–2938, San Francisco, USA, 2016.

Ahmed Ali, Stephan Vogel, and Steve Renals. Speech Recognition challenge in the Wild: Arabic MGB-3. *arXiv preprint*, arXiv:1709.07276, 2017.

Khaled Alrifai, Ghaida Rebdawi, and Nada Ghneim. Arabic Tweeps Gender and Dialect Prediction – Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

AOO Alshutayri and Eric Atwell. Exploring Twitter as a Source of an Arabic Dialect Corpus. *International Journal of Computational Linguistics (IJCL)*, 8(2):37–44, 2017.

Areej Alshutayri, Eric Atwell, AbdulRahman Alosaimy, James Dickins, Michael Ingleby, and Janet Watson. Arabic Language WEKA-Based Dialect Classifier for Arabic Automatic Speech Recognition Transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 204–211, Osaka, Japan, 2016.

Abdelmalek Amine, Zakaria Elberrichi, and Michel Simonet. Automatic Language Identification: An Alternative Unsupervised Approach Using a New Hybrid Algorithm. *International Journal of Computer Science and Applications*, 7(1):94–107, 2010.

Supriya Anand. Language Identification for Transliterated Forms of Indian Language Queries. In *Working Notes of Forum for Information Retrieval Evaluation (FIRE)*, Bangalore, India, 2014.

Stefanie Anstein. *Computational Approaches to the Comparison of Regional Variety Corpora: Prototyping a Semi-automatic System for German.* PhD thesis, University of Stuttgart, 2013.

Olga Artemenko and Margaryta Shramko. Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten. Master's thesis, Universität Hildesheim, 2005.

A. S. Babu and P. Kumar. Comparing Neural Network Approach with N-Gram Approach for Text Categorization. *International Journal on Computer Science and Engineering*, 2 (1):80–83, 2010.

Ivana Balažević, Mikio Braun, and Klaus-Robert Müller. Language Detection For Short Text Messages In Social Media. *arXiv preprint*, arXiv:1608.08515, 2016.

Timothy Baldwin and Marco Lui. Multilingual Language Identification: ALTW 2010 Shared Task Dataset. In *Proceedings of the Australasian Language Technology Workshop 2010 (ALTW 2010)*, pages 5–7, Melbourne, Australia, 2010a.

Timothy Baldwin and Marco Lui. Language Identification: The Long and the Short of the Matter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 229–237, Los Angeles, USA, 2010b.

Somnath Banerjee, Aniruddha Roy, Alapan Kuila, Sudip Kumar Naskar, Sivaji Bandyopadhyay, and Paolo Rosso. A Hybrid Approach for Transliterated Word-Level Language Identification: CRF with Post Processing Heuristics. In *Proceedings of the Sixth Workshop of the Forum for Information Retrieval Evaluation (FIRE 2014)*, pages 54–59, Bangalore, India, 2014.

Kfir Bar and Nachum 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*, pages 139–143, Doha, Qatar, 2014.

Adrien Barbaresi. An Unsupervised Morphological Criterion for Discriminating Similar Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 212–220, Osaka, Japan, 2016.

Adrien Barbaresi. Discriminating between Similar Languages using Weighted Subword Features. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 184–189, Valencia, Spain, 2017.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. Code Mixing: A Challenge for Language Identification in the Language of Social Media. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 127–132, Doha, Qatar, 2014a.

Utsab Barman, Joachim Wagner, Grzegorz Chrupala, and Jennifer Foster. DCU-UVT: Word-Level Language Classification with Code-Mixed Data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar, 2014b.

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, 2017.

Eleanor Olds Batchelder. A Learning Experience: Training an Artificial Neural Network to Discriminate Languages. Technical report, 1992.

Eda Baykan, Monika Henzinger, and Ingmar Weber. Web page language identification based on URLs. In *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB 2008)*, pages 176–187, Auckland, New Zealand, 2008.

Kenneth R. Beesley. Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text. In *Proceedings of the 29th Annual Conference of the American Translators Association: Languages at Crossroads*, pages 47–54, Seattle, USA, 1988.

Božo Bekavac, Kristina Kocijan, and Marko Tadić. Near Language Identification using NooJ. In *Formalising Natural Languages with NooJ 2014: Selected papers from the NooJ 2014 International Conference*, pages 152–166. Cambridge Scholars Publishing, Sassari, Italy, 2014.

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

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. Language Identification for Creating Language-specific Twitter Collections. In *Proceedings of the Second Workshop on Language in Social Media (LSM2012)*, pages 65–74, Montréal, Canada, 2012.

Yves Bestgen. 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*, pages 115–123, Valencia, Spain, 2017.

Aditya Bhargava and Grzegorz Kondrak. Language Identification of Names with SVMs. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 693–696, Los Angeles, California, USA, June 2010.

Rupal Bhargava, Yashvardhan Sharma, Shubham Sharma, and Abhinav Baid. Query Labeling for Indic Languages Using a Hybrid Approach. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 42–44, Gandhinagar, India, 2015.

S. Nagesh Bhattu and Vadlamani Ravi. Language Identification in Mixed Script Social Media Text. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 39–31, Gandhinagar, India, 2015.

Chris Biemann and Sven Teresniak. Disentangling from Babylonian confusion — Unsupervised Language Identification. In Alexander Gelbukh, editor, *Proceedings of the 6th*

International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005), pages 773–784, Mexico City, Mexico, 2005. Springer.

Enikö Beatrice Bilcu and Jaakko Astola. A Hybrid Neural Network for Language Identification from Text. In *Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pages 253–258, Maynooth, Ireland, 2006.

Arnold Binas. Markovian Time Series Models for Language Identification. Project Report, 2005.

Johannes Bjerva. Byte-based Language Identification with Deep Convolutional Networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 119–126, Osaka, Japan, 2016.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Su Lin Blodgett, Johnny Tian-Zheng Wei, and Brendan O'Connor. A Dataset and Classifier for Recognizing Social Media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, Copenhagen, Denmark, 2017.

Victoria Bobicev. Discriminating between Similar Languages Using PPM. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 59–65, Hissar, Bulgaria, 2015.

Gerrit Botha, Victor Zimu, and Etienne Barnard. Text-based Language Identification for South African Languages. *Transactions of the South African Institute of Electrical Engineers*, 98(4):141–148, 2007.

Gerrit Reinier Botha. Text-Based Language Identification for The South African Languages. Master's thesis, University of Pretoria, Hatfield, Pretoria, South Africa, 2008.

Gerrit Reinier Botha and Etienne Barnard. Factors that Affect the Accuracy of Text-based Language Identification. In J. R. Tapamo and F. Nicolls, editors, *Proceedings of the Eighteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 7–12, Pietermaritzburg, South Africa, 2007.

Gerrit Reinier Botha and Etienne Barnard. Factors that Affect the Accuracy of Text-based Language Identification. *Computer Speech and Language*, 26(5):307–320, October 2012.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland, 2014.

Matko Bošnjak, Eduarda Mendes Rodrigues, and Luis Sarmento. Robust Language Identification with RapidMiner: A Text Mining Use Case. In Markus Hofmann and Ralf Klinkenberg, editors, *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, pages 213–239. Chapman and Hall/CRC, 2013.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4): 467–479, 1992.

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

Ralf D. Brown. 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, 2013.

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

Brian O. Bush. Language Identification of Tweets Using LZW Compression. In *3rd Pacific Northwest Regional NLP Workshop (NW-NLP 2014)*, Redmond, USA, 2014.

Paul van Cann. Dialect Identification on Twitter: A Research About the Detection of the Limburgian Dialect from Twitter messages. Master's thesis, University of Tilburg, 2015.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic text. *Language Resources and Evaluation*, 47(1):195–215, 2013.

Dayvid Castro, Ellen Souza, and Adriano L. I. de Oliveira. Discriminating between Brazilian and European Portuguese National Varieties on Twitter Texts. In *Proceedings of the 5th Brazilian Conference on Intelligent Systems (BRACIS 2016)*, pages 265–270, Recife, Pernambuco, Brazil, 2016. IEEE.

Dayvid W. Castro, Ellen Souza, Douglas Vitório, Diego Santos, and Adriano L. I. Oliveira. Smoothed N-gram Based Models for Tweet Language Identification: A Case Study of the Brazilian and European Portuguese National Varieties. *Applied Soft Computing*, 61: 1160–1172, 2017.

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

Jordan Cazamias, Chinmayi Dixit, and Martina Marek. Large-Scale Language Classification - Writing a Detector for 200 Languages on Twitter. Stanford course report, 2015.

Çagri Çöltekin and Taraka Rama. 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, 2016.

Çagri Çöltekin and Taraka Rama. Tübingen System in VarDial 2017 Shared Task: Experiments with Language Identification and Cross-lingual Parsing. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 146–155, Valencia, Spain, 2017.

Ebru Celikel. Language Discrimination via PPM Model. In Henry Selvaraj and Pradip K. Srimani, editors, *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)*, volume 1, pages 57–62, Las Vegas, Nevada, USA, 2005.

Hakan Ceylan and Yookyung Kim. Language Identification of Search Engine Queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1066–1074, Suntec, Singapore, 2009. doi: 10.3115/1690219.1690295.

Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. Unraveling the English-Bengali Code-Mixing Phenomenon. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 80–89, Austin, TX, USA, 2016a.

Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. Columbia-Jadavpur submission for EMNLP 2016 Code-Switching Workshop Shared Task: System Description. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 112–115, Austin, TX, USA, 2016b.

Joseph Chee Chang and Chu-Cheng Lin. Recurrent-neural-network for Language Detection on Twitter Code-Switching Corpus. *arXiv preprint*, arXiv:1412.4314, 2014.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–394, 1999.

Stanley F Chen and Benoît Maison. Using Place Name Data to Train Language Identification Models. In *8th European Conference on Speech Communication and Technology EUROSPEECH 2003 - INTERSPEECH 2003*, pages 1349–1352, Geneva, Switzerland, 2003.

Tianqi Chen and Carlos Guestrin. XGBoost: Reliable Large-sclae Tree Boosting System. *arXiv preprint*, 1603.02754:1–6, 2016.

Yining Chen, Jiali You, Min Chu, Yong Zhao, and Jinlin Wang. Identifying Language Origin of Person Names with N-grams of Different Units. In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, volume 1, pages 729–732, Toulouse, France, 2006.

Yew Choong Chew, Yoshiki Mikami, Chandrajith Ashuboda Marasinghe, and S. Turrance Nandasara. Optimizing n-gram Order of an n-gram Based Language Identification Algorithm for 68 Written Languages. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 02(02):21–28, 2009.

Yew Choong Chew, Yoshiki Mikami, and Robin Lee Nagano. Language Identification of Web Pages Based on Improved N-gram Algorithm. *International Journal of Computer Science Issues*, 8(3):47–58, 2011.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India

System. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79, Doha, Qatar, 2014.

Kenneth Church. Stress Assignment in Letter to Sound Rules for Speech Synthesis. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 246–253, Chicago, Illinois, USA, 1985.

Andre Cianflone and Leila Kosseim. N-gram and Neural Language Models for Discriminating Similar Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 243–250, Osaka, Japan, 2016.

Alina Maria Ciobanu and Liviu P Dinu. A Computational Perspective on the Romanian Dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3281–3285, Portorož, Slovenia, may 2016.

Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, and Liviu P Dinu. Including Dialects and Language Varieties in Author Profiling. *arXiv preprint*, arXiv:1707.00621, 2017.

Simon Clematide and Peter Makarov. CLUZH at VarDial GDI 2017: Testing a Variety of Machine Learning Tools for the Classification of Swiss German Dialects. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 170–177, Valencia, Spain, 2017.

Michael Clyne. *Pluricentric Languages: Different Norms in Different Nations*. CRC Press, Boca Raton, USA, 1992.

Peter Constable and Gary Simons. Language identification and IT: Addressing Problems of Linguistic Diversity on a Global Scale. SIL Electronic Working Papers 2000-001, SIL International, Dallas, USA, 2000.

Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts, 1990.

Ryan Cotterell and Chris Callison-Burch. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014.

Jim Cowie, Yevgeny Ludovik, and Ron Zacharski. Language Recognition for Mono- and Multi-lingual Documents. In *Proceedings of the VexTal Conference*, pages 209–214, Venice, Italy, 1999.

Mathias Creutz. Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, volume 1, pages 280–287, Sapporo, Japan, 2003.

Marcelo Criscuolo and Sandra Maria Aluísio. Discriminating between Similar Languages with Word-level Convolutional Neural Networks. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 124–130, Valencia, Spain, 2017.

Joaquim Ferreira da Silva and Gabriel Pereira Lopes. Identification of Document Language is Not yet a Completely Solved Problem. In *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, page 212, Sydney, Australia, 2006a. IEEE.

Joaquim Ferreira da Silva and Gabriel Pereira Lopes. Identification of Document Language in Hard Contexts. In *ACM-SIGIR 2006 New Directions in Multilingual Information Access Proceedings of the Workshop*, pages 40–48, Seattle, USA, 2006b.

Joaquim Ferreira da Silva and Gabriel Pereira Lopes. Using Covariance as a Similarity Measure for Document Language Identification in Hard Contexts. *Pliska Studia Mathematica Bulgarica*, 18:341–360, 2007.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. Verifiably Effective Arabic Dialect Identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1465–1468, Doha, Qatar, 2014.

Amitava Das and Björn Gambäck. Code-Mixing in Social Media Text: The Last Language Identification Frontier? *Traitement Automatique des Langues*, 54(3):41–64, 2013.

Amitava Das and Björn Gambäck. Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 169–178, Goa, India, 2014.

Franca Debole and Fabrizio Sebastiani. An Analysis of the Relative Hardness of Reuters-21578 Subsets. *Journal of the American Society for Information Science and technology*, 56(6):584–596, 2005.

Persi Diaconis and Ronald L. Graham. Spearmam's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268, 1977.

Sascha Diwersy, Stefan Evert, and Stella Neumann. A Weakly Supervised Multivariate Approach to the Study of Language Variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. De Gruyter, Berlin, 2014.

Nina Dongen. Analysis and Prediction of Dutch-English Code-switching in Dutch Social Media Messages. Master's thesis, Universiteit van Amsterdam, Amsterdam, Netherlands, 2017.

Yerai Doval, David Vilares, and Jesús Vilares. Automatic Language Identification in Twitter: Adapting State-of-the-Art Identifiers to the Iberian Context. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 39–43, Girona, Spain, 2014.

Gabriel Doyle. Mapping Dialectal Variation by Querying Social Media. In *Proceedings of the 14th Conference of the EACL (EACL 2014)*, pages 98–106, Gothenburg, Sweden, 2014.

Ted Dunning. Statistical Identification of Language. Technical Report MCCS 940-273, Computing Research Laboratory, New Mexico State University, 1994.

Sukanya Dutta, Tista Saha, Somnath Banerjee, and Sudip Kumar Naskar. Text Normalization in Code-Mixed Social Media Text. In *2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pages 378–382, Kolkata, India, 2015.

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. Improved Text Language Identification for the South African Languages. In *Proceedings of the 28th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2017)*, Bloemfontein, South Africa, 2017.

Hisham El-Shishiny, Alexander Troussov, D. J. McCloskey, Mayo Takeuchi, Alex Nevidomsky, and Pavel Volkov. Word Fragments Based Arabic Language Identification. In *Proceedings from NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2004.

Mohamed Eldesouki, Fahim Dalvi, Hassan Sajjad, and Kareem Darwish. QCRI DSL 2016: Spoken Arabic Dialect Identification Using Textual Features. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 221–226, Osaka, Japan, 2016.

Heba Elfardy and Mona Diab. Token Level Identification of Linguistic Code Switching. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012: Posters*, pages 287–296, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Heba Elfardy and Mona Diab. Sentence Level Dialect Identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Sofia, Bulgaria, 2013.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. Code Switch Point Detection in Arabic. In E. Métais, F. Meziane, M. Sararee, V. Sugumaran, and S. Vadera, editors, *Proceedings of the Natural Language Processing and Information Systems - 18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, pages 412–416, Salford, UK, 2013. Springer.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. AIDA: Identifying Code Switching in Informal Arabic Text. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar, 2014.

Hani A. Elgabou and Dimitar Kazakov. Building Dialectal Arabic Corpora. In *The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 52–57, Varna, Bulgaria, 2017.

David Elworthy. Language Identification with Confidence Limits. In *Proceedings of the 6th Annual Workshop on Very Large Corpora*, pages 94–101, Montréal, Canada, 1998.

Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. SeedLing: Building and Using a Seed corpus for the Human Language Project.

In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–85, Baltimore, USA, June 2014. URL http://www.aclweb.org/anthology/W14-2211.

Ramy Eskander, Mohamed Al-Badrashiny, Nizar Habash, and Owen Rambow. Foreign Words and the Automatic Processing of Arabic Social Media Text Written in Roman Script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 1–12, Doha, Qatar, 2014.

Alexandra Espichán-Linares and Arturo Oncevay-Marcos. A Low-Resourced Peruvian Language Identification Model. In Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas, editors, *Proceedings of the 4th Annual International Symposium on Information Management and Big Data (SIMBig 2017)*, pages 57–63, Lima, Peru, 2017.

Alexandra Espichán-Linares and Arturo Oncevay-Marcos. Language Identification with Scarce Data: A Case Study from Peru. In *Proceedings of the 4th Annual International Symposium on Information Management and Big Data (SIMBig 2017), Revised Selected Papers*, Lima, Peru, 2018.

Raül Fabra-Boluda, Francisco Rangel, and Paolo Rosso. NLEL UPV Autoritas participation at Discrimination between Similar Languages DSL 2015 Shared Task. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 52–58, Hissar, Bulgaria, 2015.

Hector-Hugo Franco-Penya and Liliana Malmani Sanchez. Tuning Bayes Baseline for Dialect Detection. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 227–234, Osaka, Japan, 2016.

Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé, and M. Antònia Martít. Language Variety Identification Using Distributed Representations of Words and Documents. In *Proceedings of the 6th International Conference of the CLEF Association (CLEF'15): Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–40, 2015a.

Marc Franco-Salvador, Paolo Rosso, and Francisco Rangel. Distributed Representations of Words and Documents for Discriminating Similar Languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 11–16, Hissar, Bulgaria, 2015b.

Marc Franco-Salvador, Greg Kondrak, and Paolo Rosso. Bridging the Native Language and Language Variety Identification Tasks. In *Proceedings of the 21st International Conference Knowledge-Based and Intelligent Information and Engineering Systems (KES-2017)*, volume 112, pages 1554–1561, Marseille, France, 2017a. doi: https://doi.org/10.1016/j.procs.2017.08.068. URL http://www.sciencedirect.com/science/article/pii/S1877050917314126.

Marc Franco-Salvador, Nataliia Plotnikova, Neha Pawar, and Yassine Benajiba. Subword-based Deep Averaging Networks for Author Profiling – Notebook for PAN at CLEF

2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017b. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Yoav Freund and Robert E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55: 119–139, 1997.

Pablo Gamallo, Marcos Garcia, and Susana Sotelo. Comparing Ranking-based and Naive Bayes Approaches to Language Detection on Tweets. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 12–16, Girona, Spain, 2014.

Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, and Manex Agirrezabal. Comparing two Basic Methods for Discriminating Between Similar Languages and Varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 170–177, Osaka, Japan, 2016.

Pablo Gamallo, Jose Ramom Pichel, and Iñaki Alegria. A Perplexity-Based Method for Similar Languages Discrimination. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 109–114, Valencia, Spain, 2017.

Archana Garg, Vishal Gupta, and Manish Jindal. A Survey of Language Identification Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, 6 (4):388–400, 2014.

Satanu Ghosh, Souvick Ghosh, and Dipankar Das. Labeling of Query Words using Conditional Random Field. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 31–34, Gandhinagar, India, 2015.

Emmanuel Giguet. Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning. In *Proceedings of the International Workshop on Parsing Technologies (IWPT'95)*, Prague - Karlovy Vary, Czech Republic, 1995.

Emmanuel Giguet. *Méthode pour l'Analyse Automatique de Structures Formelles sur Documents Multilingues*. PhD thesis, Université de Caen, 1998.

Oluwapelumi Giwa. *Language Identification for Proper Name Pronunciation*. PhD thesis, North-West University, Vaal Triangle, 2016.

Oluwapelumi Giwa and Marelie H. Davel. N-Gram based Language Identification of Individual Words. In Philip Robinson, editor, *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 15–22, Johannesburg, South Africa, 2013.

Oluwapelumi Giwa and Marelie H. Davel. Language Identification of Individual Words with Joint Sequence Models. In *Proceedings of Interspeech 2014*, Singapore, 2014.

Moises Goldszmidt, Marc Najork, and Stelios Paparizos. Boot-Strapping Language Identifiers for Short Colloquial Postings. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2013), Part II*, pages 95–111, Prague, Czech Republic, 2013. Springer.

Thomas Gottron and Nedim Lipka. A Comparison of Language Identification Approaches on Short, Query-style Texts. In *Advances in Information Retrieval - Proceedings of the 32nd annual European Conference on Information Retrieval Research (ECIR 2010)*, pages 611–614, Milton Keynes, UK, 2010. Springer.

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

Cyril Goutte and Serge Léger. Advances in Ngram-based Discrimination of Similar Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–184, Osaka, Japan, 2016.

Cyril Goutte, Serge Léger, and Marine Carpuat. The NRC System for Discriminating Similar Languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 139–145, Dublin, Ireland, 2014.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may 2016.

Gregory Grefenstette. Comparing Two Language Identification Schemes. In *Proceedings of the 3rd International conference on Statistical Analysis of Textual Data (JADT 1995)*, Rome, Italy, 1995.

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

Imène Guellil and Faiçal Azouaou. Arabic Dialect Identification With an Unsupervised Learning (based on a lexicon). Application case: Algerian Dialect. In *Proceedings of the 2016 IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (CSE-EUC-DCABES 2016)*, pages 724–731, Paris, France, 2016.

Bhumika Gupta, Gaurav Bhatt, and Ankush Mittal. Language Identification and Disambiguation in Indian Mixed-Script. In Nikolaj Bjørner, Sanjiva Prasad, and Laxmi Parida, editors, *Distributed Computing and Internet Technology*, pages 113–121. Springer, 2016.

Deepak Kumar Gupta, Shubham Kumar, and Asif Ekbal. Machine Learning Approach for Language Identification & Transliteration: Shared Task Report of IITP-TS. In *Forum for Information Retrieval Evaluation (FIRE)*, pages 60–64, Bangalore, India, 2014.

Gualberto A. Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara Bullock, and Almeida Jacqueline Toribio. Moving Code-switching Research Toward More Empirically Grounded Methods. In Thierry Declerck and Sandra Kübler, editors, *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*, pages 1–9, Bloomington, IN, USA, 2017.

Helena Gómez-Adorno, Ilia Markov, Jorge Baptista, Grigori Sidorov, and David Pinto. Discriminating between Similar Languages Using a Combination of Typed and Untyped Character N-grams and Words. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 137–145, Valencia, Spain, 2017.

Nizar Habash and Fatiha Sadat. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the 7th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL06)*, pages 49–52, New York, NY, USA, 2006.

Andrea Haţegan, Bogdan Bârligă, and Ioan Tăbuş. Language Identification of Individual Words in a Multilingual Automatic Speech Recognition System. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 4357–4360, Taipei, Taiwan, 2009. IEEE.

Harald Hammarström. A Fine-Grained Model for Language Identification. In *Proceedings of Improving Non English Web Searching (iNEWS-07) Workshop at SIGIR 2007*, pages 14–20, Amsterdam, Netherlands, 2007.

Amir Hamzah. Deteksi bahasa untuk dokumen teks berbahasa Indonesia. In *Seminar Nasional Informatika 2010 (semnasIF 2010)*, pages A5–A13, Jakarta, Indonesia, 2010.

Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor. Classifying ASR Transcriptions According to Arabic Dialect. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 126–134, Osaka, Japan, 2016.

Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor. Identifying Dialects with Textual and Acoustic Cues. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–101, Valencia, Spain, 2017.

Farheen Hanif, Fouzia Latif, and M. Sikandar Hayat Khiyal. Unicode Aided Language Identification across Multiple Scripts and Heterogeneous Data. *Information Technology Journal*, 6(4):534–540, 2007.

Maimaitiyiming Hasimu and Wushouer Silamu. Three-stage Short Text Language Identification Algorithm. *Journal of Digital Information Management*, 15(6):354–371, December 2017.

Katia Hayati. Language Identification on the World Wide Web. Master's thesis, University of California Santa Cruz, Santa Cruz, California, USA, 2004.

Sengül Bayrak Hayta, Hidayet Takçı, and Mübariz Eminli. Language Identification Based On N-Gram Feature Extraction Method By Using Classifiers. *IU-Journal of Electrical and Electronics Engineering*, 13(2):1629–1639, 2013.

Junqing He, Zhen Zhang, Xuemin Zhao, Peijia Li, and Yonghong Yan. Similar Language Identification for Uyghur and Kazakh on Short Spoken Texts. In *Proceedings of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC 2016)*, volume 2, pages 496–499, Hangzhou, China, 2016.

Robert Hecht-Nielsen. Theory of the Backpropagation Neural Network. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 1989)*, pages I593–I605, Washington, DC, USA, 1989. doi: 10.1109/IJCNN.1989.118638.

Peter Henrich. Language Identification for the Automatic Grapheme-to-phoneme Conversion of Foreign Words in a German Text-to-speech System. In *First European Conference on Speech Communication and Technology*, pages 2220–2223, Paris, France, 1989.

Ondřej Herman, Vít Suchomel, Vít Baisa, and Pavel Rychlý. DSL Shared task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation-Maximization and Chunk-based Language Model. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–118, Osaka, Japan, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9:1735–1780, 1997.

Nora Hollenstein and Noëmi Aepli. A Resource for Natural Language Processing of Swiss German Dialects. In *Proceedings of GSCL*, pages 108–109, 2015.

Arthur S. House and Edward P. Neuburg. Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations. *The Journal of the Acoustical Society of America*, 62(3):708–713, 1977.

Chu-Ren Huang and Lung-Hao Lee. Contrastive Approach towards Text Source Classification based on Top-Bag-of-Word Similarity. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 404–410, Cebu City, Philippines, November 2008.

Fei Huang. Improved Arabic Dialect Classification with Social Media Data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 2118–2126, Lisbon, Portugal, 2015. URL `https://aclweb.org/anthology/D/D15/D15-1254`.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. Reconsidering Language Identification for Written Language Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 485–488, Genoa, Italy, 2006.

Lluís-F. Hurtado, Ferran Pla, Mayte Giménez, and Emilio Sanchis. ELiRF-UPV at TweetLID: Twitter Language Identification. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 35–38, Girona, Spain, 2014.

Juha Häkkinen and Jilei Tian. N-gram and Decision Tree Based Language Identification for Written Words. In *Conference Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, pages 335–338, Madonna di Campiglio, Italy, 2001.

K. Indhuja, M. Indu, C. Sreejith, and P. C. Reghu Raj. Text Based Language Identification System for Indian Languages Following Devanagiri Script. *International Journal of Engineering Reseach and Technology*, 3(4):327–331, 2014.

Radu Tudor Ionescu. Local rank distance. In Nikolaj Björner, Viorel Negru, Tetsuo Ida, Tudor Jebelean, Dana Petcu, Stephen Watt, and Daniela Zaharie, editors, *Proceedings of the 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2013)*, pages 219–226, Timisoara, Romania, 2013.

Radu Tudor Ionescu and Andrei M. Butnaru. Learning to Identify Arabic and German Dialects using Multiple Kernels. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 200–209, Valencia, Spain, 2017.

Radu Tudor Ionescu and Marius Popescu. UnibucKernel: An Approach for Arabic Dialect Identification based on Multiple String Kernels. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 135–144, Osaka, Japan, 2016.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. Hierarchical Character-Word Models for Language Identification. In *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 84–93, Austin, TX, USA, 2016a.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. A Neural Model for Language Identification in Code-Switched Tweets. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 60–64, Austin, TX, USA, 2016b.

Devanshu Jain. DA-IICT in FIRE 2015 Shared Task on Mixed Script Information Retrieval. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 53–56, Gandhinagar, India, 2015.

Radwan Jalam. *Apprentissage Automatique et Catégorisation de Textes Multilingues*. PhD thesis, Université Lumière Lyon 2, 2003.

Radwan Jalam and Olivier Teytaud. Kernel-based Text Categorization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'01)*, volume 3, pages 1891–1896, Washington, DC, USA, 2001a.

Radwan Jalam and Olivier Teytaud. Identification de la Langue et Catégorisation de Textes Basées sur les N-grammes. In Henri Briand and Fabrice Guillet, editors, *Journées Francophones d'extraction et de gestion de connaissances (EGC'2001)*, pages 227–238, Nantes, France, 2001b.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. The Finno-Ugric Languages and The Internet Project. *Septentrio Conference Series*, 0(2):87–98, 2015a. ISSN 2387-3086. doi: 10.7557/5.3471. URL http://septentrio.uit.no/index.php/SCS/article/view/3471.

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

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. Discriminating Similar Languages with Token-based Backoff. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, LT4VarDial '15*, pages 44–51, Hissar, Bulgaria, 2015b.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 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, 2015c.

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

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

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

Harsh Jhamtani, Suleep Kumar Bhogi, and Vaskar Raychoudhury. Word-level Language Identification in Bi-lingual Code-switched Texts. In *28th Pacific Asia Conference on Language, Information and Computation*, pages 348–357, Phuket, Thailand, 2014.

Jay J. Jiang and David W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference on Research on Computational Linguistics (ROCLING)*, Taiwan, 1997.

Taeho Jo. Neural Text Categorizer for Exclusive Text Categorization. *Journal of Infomration Processing Systems*, 4:77–86, 2008.

Hardik Joshi, Apurva Bhatt, and Honey Patel. Transliterated Search using Syllabification Approach. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2013)*, New Delhi, India, 2013.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, volume 2, pages 427–431, Valencia, Spain, 2017.

Pierre Jourlin. Entity Recognition and Language Identification with FELTS. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

P. Juola. Language Identification, Automatic. In *Encyclopedia of Language and Linguistics*, volume 6, pages 508—-510. Elsevier, Amsterdam, Netherlands, 2006.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2, pages 51–57, Vancouver, Canada, 2017.

Said Kadri and Abdelouahab Moussaoui. An Effective Method to Recognize the Language of a Text in a Collection of Multilingual Documents. In *Proceedings of the International Conference on Electronics, Computer and Computation (ICECCO 2013)*, pages 208–211, Ankara, Turkey, 2013.

Slava M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(3):400–401, March 1987.

Thomas Kerwin. Classification of Natural Language Based on Character Frequency. Ohio Supercomputer Center, 2006.

Guillaume Kheng, Laporte Léa, and Michael Granitzer. INSA LYON and UNI PASSAU's participation at PAN@CLEF'17: Author Profiling task - Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Gen-Itiro Kikui. Identifying the Coding System and Language of Online Documents on the Internet. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pages 652–657, Copenhagen, Denmark, 1996.

Seungbeom Kim and Jongsoo Park. Automatic Detection of Character Encoding and Language. Technical report, Stanford University, 2007.

Ben King and Steven Abney. 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, June 2013.

Ben King, Dragomir Radev, and Steven Abney. Experiments in Sentence Language Identification with Groups of Similar Languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Dublin, Ireland, 2014a.

Josh King and Jon Dehdari. An N-gram Based Language Identification System. The Ohio State University, 2008.

Levi King, Eric Baucom, Timur Gilmanov, Sandra Kübler, Daniel Whyatt, Wolfgang Maier, and Paul Rodrigues. The IUCL+ System: Word-Level Language Identification via Extended Markov Models. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 102–106, Doha, Qatar, 2014b.

Levi King, Sandra Kübler, and Wallace Hooper. Word-level language identification in The Chymistry of Isaac Newton. *Digital Scholarship in the Humanities*, 30(4):532–540, 2015.

Tom Kocmi and Ondřej Bojar. LanideNN: Multilingual Language Identification on Character Window. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Long Papers*, volume 1, pages 927–936, Valencia, Spain, 2017.

Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak. Author Profiling with Bidirectional RNNs using Attention with GRUs - Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Stasinos Konstantopoulos. What's in a Name? In *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP-07)*, Borovets, Bulgaria, 2007.

Anett Kralisch and Thomas Mandl. Barriers to Information Access Across Languages on the Internet: Network and Language Effects. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, volume 3, page 54b, Kauai, USA, 2006.

Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. Language identification based on string kernels. In *Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT-2005)*, volume 2, pages 896–899, Beijing, China, 2005.

Stan Kulikowski. Language Identification of Short Texts. Newsgroup article, Educational Research and Development Center, The University of West Florida, 1991.

Rahul Venkatesh RM Kumar, Anand M Kumar, and KP Soman. AmritaCEN_NLP @ FIRE 2015 Language Identification for Indian Languages in Social Media Text. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 28–30, Gandhinagar, India, 2015.

Gustavo Laboreiro, Matko Bošnjak, Luís Sarmento, Eduarda Mendes Rodrigues, and Eugénio Oliveira. Determining Language Variant in Microblog Messages. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC'13)*, pages 902–907, Coimbra, Portugal, 2013.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL `http://dl.acm.org/citation.cfm?id=645530.655813`.

Priyadarshini Lamabam and Kunal Chakma. A Language Identification System for Code-Mixed English-Manipuri Social Media Text. In *Proceedings of the IEEE International Conference on Engineering and Technology (ICETECH 2016)*, pages 79–83, Coimbatore, Tamil Nadu, India, 2016.

Stefan Langer. Natural Languages and the World Wide Web. *Bulletin de Linguistique Appliquée et Générale (BULAG)*, 26:89–100, 2001.

Sebastian Leidig. Single and Combined Features for the Detection of Anglicisms in German and Afrikaans. Bachelor's Thesis, Karlsruhe Institute of Technology, 2014.

William D Lewis and Fei Xia. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *Literary and Linguistic Computing*, 25(3):303–319, 2010.

Shanjian Li and Katsuhiko Momoi. A Composite Approach to Language/Encoding Detection. In *Nineteenth International Unicode Conference (IUC19)*, San Jose, California, USA, 2001.

Yitong Li, Trevor Cohn, and Timothy Baldwin. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2018)*, pages 474–479, New Orleans, USA, 2018.

Chu-Cheng Lin, Waleed Ammar, Lori Levin, and Chris Dyer. The CMU Submission for the Shared Task on Language Identification in Code-Switched Data. In *Proceedings of First Workshop on Computational Approaches to Code Switching*, pages 80–86, Doha, Qatar, 2014.

Dekang Lin. An Information-Theoretic Definition of Similarity. In Jude W. Shavlik, editor, *Proceedings of the 15th International Conference on Machine Learning (ICML-98)*, pages 296–304, Madison, Wisconsin, USA, 1998.

Yoseph Linde, Andrés Buzo, and Robert M. Gray. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, COM-28(1):84–95, 1980.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. Microblogs as Parallel Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria, August 2013.

Rafael Dueire Lins and Paulo Jr. Gonçalves. Automatic Language Identification of Written Texts. In *Proceedings of the 2004 ACM symposium on Applied Computing (SAC 2004)*, pages 1128–1133, Nicosia, Cyprus, 2004.

Nick Littlestone. Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. *Machine learning*, 2(4):285–318, 1987.

Nikola Ljubešić and Denis Kranjcić. Discriminating between VERY Similar Languages among Twitter Users. In *Proceedings of the 9th Language Technologies Conference*, pages 90–94, Ljubljana, Slovenia, 2014.

Nikola Ljubešić and Denis Kranjcić. Discriminating Between Closely Related Languages on Twitter. *Informatica*, 39, 2015.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. Language Identification: How to Distinguish Similar Languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007)*, pages 541–546, Cavtat/Dubrovnik, Croatia, 2007.

Man Lu and Moustafa Mohamed. LAHGA: Arabic Dialect Classifier. Unpublished Report. December 2011.

Yevgeny Ludovik and Ron Zacharski. Multilingual Document Language Recognition for Creating Corpora. Technical report, New Mexico State University, 1999.

Marco Lui. *Generalized Language Identification*. PhD thesis, The University of Melbourne, 2014.

Marco Lui and Timothy Baldwin. Cross-domain Feature Selection for Language Identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, volume 1, pages 553—-561, Chiang Mai, Thailand, 2011.

Marco Lui and Timothy Baldwin. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea, 2012.

Marco Lui and Timothy Baldwin. Accurate Language Identification of Twitter Messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Marco Lui and Paul Cook. Classifying English Documents by National Dialect. In *Proceedings of Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, Brisbane, Australia, December 2013.

Marco Lui, Jey Han Lau, and Timothy Baldwin. Automatic Detection and Language Identification of Multilingual Documents. *Transactions of the Association for Computational Linguistics*, 2:27–40, 2014a. ISSN 2307-387X. URL `https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/86`.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. Exploring Methods and Resources for Discriminating Similar Languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 129–138, Dublin, Ireland, August 2014b.

Shane MacNamara, Pádraig Cunningham, and John Byrne. Neural Networks for Language Identification: a Comparative Study. *Information processing and management*, 34(4): 395–403, 1998.

Wolfgang Maier and Carlos Gómez-Rodríguez. Language Variety Identification in Spanish Tweets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants (LT4CloseLang 2014)*, pages 25–35, Doha, Qatar, October 2014. Association for Computational Linguistics.

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

Martin Majliš. Yet Another Language Identifier. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 46–54, Avignon, France, 2012.

Shervin Malmasi. Open-Set Language Identification. *arXiv preprint*, arXiv:1707.04817, 2017.

Shervin Malmasi and Mark Dras. Automatic Language Identification for Persian and Dari Texts. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics, PACLING'15*, pages 59–64, Bali, Indonesia, 2015a.

Shervin Malmasi and Mark Dras. Language Identification using Classifier Ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, LT4VarDial'15*, pages 35–43, Hissar, Bulgaria, 2015b.

Shervin Malmasi and Mark Dras. Feature Hashing for Language and Dialect Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 399–403, Vancouver, Canada, 2017.

Shervin Malmasi and Marcos Zampieri. Arabic Dialect Identification in Speech Transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 106–113, Osaka, Japan, 2016.

Shervin Malmasi and Marcos Zampieri. German Dialect Identification in Interview Transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 164–169, Valencia, Spain, 2017a.

Shervin Malmasi and Marcos Zampieri. Arabic Dialect Identification Using iVectors and ASR Transcripts. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–183, Valencia, Spain, 2017b.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics, PACLING'15*, pages 209–217, Bali, Indonesia, 2015.

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

Soumik Mandal, Somnath Banerjee, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. Adaptive Voting in Multiple Classifier Systems for Word Level Language Identification. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 49–52, Gandhinagar, India, 2015.

Thomas Mandl, Margaryta Shramko, Olga Tartakovski, and Christa Womser-Hacker. Language Identification in Multi-lingual Web-Documents. In *Proceedings of the 11th International Conference on Applications of Natural Language to Information Systems (NLDB 2006)*, pages 153–163, Klagenfurt, Austria, 2006.

Jean-Christophe Marcadet, Volker Fischer, and Claire Waast-Richard. A Transformation-based Learning Approach to Language Identification for Mixed-lingual Text-to-speech Synthesis. In *Proceedings of the 6th Interspeech and 9th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2248–2251, Lisbon, Portugal, 2005.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1993.

Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling—Notebook for PAN at CLEF 2017. In *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, 2017.

Puji Martadinata, Bayu Distiawan Trisedya, Hisar Maruli Manurung, and Mirna Adriani. Building Indonesian Local Language Detection Tools Using Wikipedia Data. In Yohei Murakami and Donghui Lin, editors, *Worldwide Language Service Infrastructure*, pages 113–123. Springer, 2016.

Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. PAN 2017: Author Profiling - Gender and Language Variety Prediction—Notebook for PAN at CLEF 2017. In *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, 2017.

Bruno Martins and Mário J. Silva. Language Identification in Web Pages. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 764–768, Santa Fe, USA, 2005.

Laura A. Mather. A Linear Algebra Approach to Language Identification. In E. V. Munson, editor, *Proceedings of the 4th International Workshop Principles of Digital Document Processing (PODDP'98)*, pages 92–103, Saint Malo, France, 1998.

Priyank Mathur, Arkajyoti Misra, and Emrah Budur. LIDE: Language Identification from Text Documents. *arXiv preprint*, arXiv:1701.03682, 2017.

Uwe F. Mayer. Bootstrapped Language Identification for Multi-site Internet Domains. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012)*, pages 579–585, Beijing, China, 2012.

Paul McNamee. Language Identification: A Solved Problem Suitable for Undergraduate Instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101, 2005.

Paul McNamee. Language and Dialect Discrimination Using Compression-Inspired Language Models. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 195–203, Osaka, Japan, 2016.

Maria Medvedeva, Martin Kroon, and Barbara Plank. When Sparse Traditional Models Outperform Dense Neural Networks: the Curious Case of Discriminating between Similar Languages. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 156–163, Valencia, Spain, 2017.

Iosu Mendizabal, Jeroni Carandell, and Daniel Horowitz. TweetSafa: Tweet Language Identification. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 21–25, Girona, Spain, 2014.

Ismael Mendoza and Julia Mendelsohn. Exploring Techniques in Distinguishing Similar Languages. Stanford course project, 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Dean Jeffrey. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL], 2013.

Rachel Mary Milne, Richard A. O'Keefe, and Andrew Trotman. A study in Language Identification. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, pages 88–95, Dunedin, New Zealand, 2012.

Akshay Minocha and Francis M. Tyers. Subsegmental Language Detection in Celtic Language Text. In *Proceedings of the First Celtic Language Techonology Workshop (CLTW 2014)*, pages 76–80, Dublin, Ireland, 2014.

Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, Shotaro Misawa, and Tomoko Ohkuma. Using Social Networks to Improve Language Variety Identification with Neural Networks. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 263–270, Taipei, Taiwan, 2017.

Serguei A. Mokhov. Complete Complimentary Results Report of the MARF's NLP Approach to the DEFT 2010 Competition. *CoRR*, abs/1006.3787, 2010a.

Serguei A Mokhov. A MARF Approach to DEFT 2010. In *Proceedings of the 6th DEFT Workshop (DEFT'10)*, pages 35–49, 2010b.

Giovanni Molina, Nicolas Rey-Villamizar, Thamar Solorio, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, and Mona Diab. 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*, pages 40–49, Austin, TX, USA, 2016.

Avashlin Moodley. Language Identification With Decision Trees: Identification Of Individual Words In The South African Languages. Bachelor's Thesis, University of South Africa, 2016.

Abhinav Mukherjee, Anirudh Ravi, and Kaustav Datta. Mixed-script Query Labelling Using Supervised Learning and Ad Hoc Retrieval Using Sub Word Indexing. In *FIRE '14 Proceedings of the Forum for Information Retrieval*, pages 86–90, Bangalore, India, 2014.

Kavi Narayana Murthy and G. Bharadwaja Kumar. Language Identification from Small Text Samples. *Journal of Quantitative Linguistics*, 13(1):57–80, January 2006.

Seppo Mustonen. Multiple Discriminant Analysis in Linguistic Problems. *Statistical Methods in Linguistics*, 4:37–44, 1965.

Yeshwant K Muthusamy and A Lawrence Spitz. Automatic Language Identification. In Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, Antonio Zampolli, and Victor Zue, editors, *Web Edition: Survey of the State of the Art in Human Language Technology*, pages 314–317. Cambridge University Press, Cambridge, UK, 1997.

Yukio Nakamura. Identification of Languages with Short Sample Texts – A Linguometric Study. *Library and information science*, 9:459–481, 1971.

Choon-Ching Ng and Ali Selamat. Improved Letter Weighting Feature Selection on Arabic Script Language Identification. In *Proceedings of the 1st Asian Conference on Intelligent Information and Database Systems (ACIIDS 2009)*, pages 150–154, Dong Hoi, Vietnam, 2009. IEEE.

Dong Nguyen and Leonie Cornips. Automatic Detection of Intra-Word Code-Switching. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86, Berlin, Germany, 2016.

Dong Nguyen and A. Seza Dogruöz. Word Level Language Identification in Online Multilingual Communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 857–862, Seattle, USA, 2013.

Shiho Nobesawa and Ikuo Tahara. Language Identification for Person Names Based on Statistical Information. In *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation (PACLIC 19)*, number 1027 in Y05, Taipei, Taiwan, December 2005.

Norzaidah Md Noh, Mohd Rusydi Abdul Talib, Azlin Ahmad, Shamimi A. Halim, and Azlinah Mohamed. Malay Language Document Identification Using BPNN. In Nikos E. Mastorakis, Anca Croitoru, Valentina Emilia Balas, Eduard Son, and Valeri Mladenov, editors, *Proceedings of the 10th WSEAS international conference on Neural networks*, pages 163–168, Prague, Czech Republic, 2009.

Daisuke Okanohara and Jun'ichi Tsujii. Text Categorization with All Substring Features. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 838–846, Miami, USA, 2009.

Rodrigo Ribeiro Oliveira and Rosalvo Ferreira de Oliveira Neto. Using Character N-grams and Style Features for Gender and Language Variety Classification – Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Gorkem Ozbek, Itamar Rosenn, and Eric Yeh. Language Classification in Multilingual Documents. Technical report, Stanford University, 2006.

Leonid Panich. Comparison of Language Identification Techniques. Bachelor's Thesis, Heinrich Heine Universität Düsseldorf, 2015.

Evangelos E. Papalexakis, Dong Nguyen, and A. Seza Dogruöz. Predicting Code-Switching in Multilingual Communication for Immigrant Communities. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 42–50, Doha, Qatar, 2014.

Kosuru Pavan, Niket Tandon, and Vasudeva Varma. Addressing Challenges in Automatic Language Identification of Romanized Text. In *Proceedings of 8th International Conference on Natural Language Processing (ICON-2010)*, Kharagpur, India, 2010.

Timo Pawelka and Elmar Jürgens. Is This Code Written in English? A Study of the Natural Language of Comments and Identifiers in Practice. In *Proceedings of the 31st International Conference on Software Maintenance and Evolution (ICSME)*, Bremen, Germany, 2015.

Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. The Automatic Identification of Lexical Variation Between Language Varieties. *Natural Language Engineering*, 16(04):469–491, October 2010.

Fuchun Peng and Dale Schuurmans. Combining Naive Bayes and n-Gram Language Models for Text Classification. In *Proceedings of the 25th European Conference on IR Research, Advances in Information Retrieval: (ECIR 2003)*, pages 335–350, Pisa, Italy, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-36618-8. doi: 10.1007/3-540-36618-0_24. URL `http://dx.doi.org/10.1007/3-540-36618-0_24`.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese Segmentation and New Word Detection Using Conditional Random Fields. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 562–568, Geneva, Switzerland, 2004.

Gergely Pethö and Eszter Mózes. An N-gram-based Language Identification Algorithm for Variable-length and Variable-language Texts. *Argumentum*, 10:56–82, 2014.

Tuan Pham and Dat Tran. VQ-based Written Language Identification. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications (ISSPA 2003)*, volume 1, pages 513–516, Paris, France, 2003.

Wikus Pienaar and Dirk Snyman. Spelling Checker-based Language Identification for the Eleven Official South African Languages. In *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa*, pages 213–217, Stellenbosch, South Africa, 2010.

Mario Piergallini, Rouzbeh Shirvani, Gauri S. Gautam, and Mohamed Chouikha. Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 21–29, Austin, TX, USA, 2016a.

Mario Piergallini, Rouzbeh Shirvani, Gauri S. Gautam, and Mohamed 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*, pages 116–120, Austin, TX, USA, 2016b.

Ferran Pla and Lluís-F. Hurtado. Language Identification in Twitter: A Study Case of Multiclass and Multilabel Text Classification Problem. *International Journal of Computational Linguistics and Applications*, 6(1):135–150, 2015.

Ferran Pla and Lluís-F. Hurtado. Language Identification of Multilingual Posts from Twitter: A Case Study. *Knowledge and Information Systems*, 51(3):965–989, 2017.

Sergi Plaza Cagigós. Catdetect, a framework for detecting Catalan tweets. Bachelor thesis, Universitat de Lleida, 2017.

Jordi Porta. Twitter Language Identification using Rational Kernels and its Potential Application to Sociolinguistics. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 17–20, Girona, Spain, 2014.

Jordi Porta and José-Luis Sancho. Using Maximum Entropy Models to Discriminate between Similar Languages and Varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 120–128, Dublin, Ireland, 2014.

Adam Poulston, Zeerak Waseem, and Mark Stevenson. Using TF-ID n-gram and Word Embedding Cluster Ensembles for Author Profiling – Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Arjen Poutsma. Applying Monte Carlo Techniques to Language Identification. *Language and Computers*, 45(1):179–189, 2002.

John 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.

M. A. Nejla Qafmolla. Automatic Language Identification. *European Journal of Language and Literature Studies*, 7(1):140–150, 2017.

Shenglan Qiao and Daniel Lévy. Similar Language Detection. Stanford University, 2015.

Will Radford and Matthias Gallé. Discriminating Between Similar Languages in Twitter Using Label Propagation. *arXiv preprint*, arXiv:1607.05408, 2016.

K. A. Rafidha Rehiman, A. S. Keerthy, K. S. Lakshmi, and A. Sreekumar. A Language Identification and Conversion System for Malayalam to Ensure Security. In *3rd National Conference on Indian Language Computing (NCILC 2013)*, Cochin, Kerala, India, 2013.

Khyathi C. Raghavi, Manoj Chinnakotla, and Manish Shrivastava. "Answer ka type kya he?" Learning to Classify Questions in Code-Mixed Language. In *WWW'15 Companion Proceedings of the 24th International Conference on World Wide Web*, pages 853–858, Florence, Italy, 2015.

Abhinav Raj and Sankha Karfa. A List-searching Based Approach for Language Identification in Bilingual Text: Shared Task Report by Asterisk. In *Working Notes of the Shared Task on Transliterated Search at Forum for Information Retrieval Evaluation FIRE '14*, Bangalore, India, 2014.

Carlos Ramisch. N-gram Models for Language Detection. Technical Report, 2008.

Bali Ranaivo-Malançon. Automatic Identification of Close Languages – Case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2(2): 126–134, 2006.

Bali Ranaivo-Malançon and Pek Kuan Ng. Language Identifier for Bahasa Malaysia and Bahasa Indonesia. In *Proceedings of the 1st Malaysian Software Engineering Conference (MySEC'05)*, pages 257–259, Penang, Maylaysia, 2005.

Francisco Rangel, Marc Franco-Salvador, and Paolo Rosso. A Low Dimensionality Representation for Language Variety Identification. In *Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, 2017a.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017b. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Prakash Ranjan, Bharathi Raja, Ruba Priyadharshini, and Rakesh Chandra Balabantaray. A Comparative Study on Code-Mixed Data of Indian Social Media vs Formal Text. In S. K. Niranjan and V. N. Manjunatha Aradhya, editors, *Proceedings of the 2nd International Conference on Contemporary Computing and Informatics (IC3I 2016)*, pages 608–611, Noida, India, 2016. IEEE.

Morton David Rau. Language Identification by Statistical Analysis. Master's thesis, Naval Postgraduate School, Monterey, 1974.

Radim Rehůřek and Milan Kolkus. Language Identification on the Web: Extending the Dictionary Method. In *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009)*, pages 357–368, Mexico City, Mexico, 2009.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62: 107–136, 2006.

Shruti Rijhwani, Royal Sequeira, Monojit Choudhury, Kalika Bali, and Chandra Sekhar Maddila. Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1971–1982, Vancouver, Canada, 2017.

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

Harald Romsdorfer and Beat Pfister. Text Analysis and Language Identification for Polyglot Text-to-speech Synthesis. *Speech communication*, 49(9):697–724, 2007.

Mike Rosner and Paulseph-John Farrugia. A Tagging Algorithm for Mixed Language Identification in a Noisy Domain. In *INTERSPEECH-2007, 8th Annual Conference of the International Speech Communication Association*, pages 190–193, Antwerp, Belgium, 2007.

Neil C. Rowe, Riqui Schwamm, and Simson L Garfinkel. Language Translation for File Paths. *Digital Investigation*, 10, 2013.

Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. Overview of the FIRE 2013 Track on Transliterated Search. In Prasenjit Majumder, Mandar Mitra, Maghulika Agrawal, and Parth Mehta, editors, *Proceedings of the 5th Forum on Information Retrieval Evaluation (FIRE '13)*, New Delhi, India, 2013. ACM.

Graham Russell and Guy Lapalme. Automatic Identification of Language and Encoding. Technical report, Laboratoire de Recherche Appliquée en Linguistique Informatique (RALI), Université de Montréal, 2003.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. Automatic Identification of Arabic Language Varieties and Dialects in Social Media. In Shou-de Lin, Lun-Wei Ku, Erik Cambria, and Tsung-Ting Kuo, editors, *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP 2014)*, pages 22–27, Dublin, Ireland, 2014a.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the first international workshop on Social media retrieval and analysis (SoMeRA 2014)*, pages 35–40, Gold Coast, QLD, Australia, 2014b. ACM.

Navanath Saharia. Phone-based Identification of Language in Code-mixed Social Network Data. *Journal of Statistics and Management Systems*, 20(4):565–574, 2017.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. Sentence Level Dialect Identification for Machine Translation System Selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778, Baltimore, USA, 2014.

Younes Samih. *Dialectal Arabic Processing Using Deep Learning*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany, 2017.

Younes Samih and Wolfgang Maier. Detecting Code-Switching in Moroccan Arabic Social Media. In *Proceedings of the 4th International Workshop on Natural Language Processing for Social Media (SocialNLP 2016 IJCAI)*, New York City, USA, 2016.

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. Multilingual Code-switching Identification via LSTM Recurrent Neural Networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, TX, USA, 2016.

Miguel A. Sanchez-Perez, Ilia Markov, Helena Gómez-Adorno, and Sidorov Grigori. Comparison of Character N-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus. In Gareth J. F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Proceedings of the 8th International Conference of the CLEF Association (CLEF 2017)*, pages 145–151, Dublin, Ireland, 2017. Springer.

Kevin P Scannell. The Crúbadán Project: Corpus Building for Under-resourced Languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, pages 5–15, Louvain-la-Neuve, Belgium, 2007.

Nils Schaetti. UniNE at CLEF 2017: TF-IDF and Deep-Learning for Author Profiling - Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Yves Scherrer and Owen Rambow. Word-based Dialect Identification with Georeferenced Rules. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1151–1161, Massachusetts, USA, 2010. Association for Computational Linguistics.

Sarah Schulz and Mareike Keller. Code-Switching Ubique Est - Language Identification and Part-of-Speech Tagging for Historical Mixed Text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, pages 43–51, Berlin, Germany, 2016.

Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

Frank Seifart and Roger Mundry. Quantitative Comparative Linguistics based on Tiny Corpora: N-gram Language Identification of Wordlists of Known and Unknown Languages from Amazonia and Beyond. *Journal of Quantitative Linguistics*, 22(3), 2015.

Ali Selamat and Nicholas Akosu. Word-length Algorithm for Language Identification of Under-resourced Languages. *Journal of King Saud University - Computer and Information Sciences*, 28:457–469, 2016.

Ali Selamat and Choon-Ching Ng. Arabic Script Documents Language Identifications Using Fuzzy ART. In David Al-Dabass, Steve Turner, Gary Tan, and Ajith Abraham, editors, *Proceedings of the Second Asia International Conference on Modeling & Simulation (AMS 2008)*, pages 528–533, Kuala Lumpur, Malaysia, 2008.

Ali Selamat, Choon-Ching Ng, and Yoshiki Mikami. Arabic Script Web Documents Language Identification Using Decision Tree-ARTMAP Model. In *Proceedings of the International Conference on Convergence Information Technology (ICCIT 2007)*, pages 717–722, Gyeongju, Korea, 2007. IEEE.

Royal Sequeira, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, and Kunal Chakma. Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 21–27, Gandhinagar, India, 2015.

Saatvik Shah, Vaibhav Jain, Sarthak Jain, Anshul Mittal, Jatin Verma, Shubham Tripathi, and Dr. Rajesh Kumar. Hierarchical classification for Multilingual Language Identification and Named Entity Recognition. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2015)*, pages 21–27, Gandhinagar, India, 2015.

H. L. Shashirekha. Automatic Language Identification from Written Texts - An Overview. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(5):156–160, 2014.

Karen Shiells and Peter Pham. Unsupervised Clustering for Language Identification. Project Report, Stanford University, 2010.

Prajwol Shrestha. Incremental N-gram Approach for Language Identification in Code-Switched Text. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 133–138, Doha, Qatar, 2014.

Prajwol Shrestha. Codeswitching Detection via Lexical Features using Conditional Random Fields. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 121–126, Austin, TX, USA, 2016.

Penelope Sibun and Jeffrey C. Reynar. Language Identification: Examining the Issues. In *Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval (SDAIR-96)*, pages 125–135, Las Vegas, USA, 1996.

Sebastian Sierra, Manuel Montes-y Gómez, Thamar Solorio, and Fabio A. González. Convolutional Neural Networks for Author Profiling—Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Utpal Kumar Sikdar and Björn Gambäck. Language Identification in Code-Switched Text Using Conditional Random Fields and Babelnet. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 127–131, Austin, TX, USA, 2016.

Vasiliki Simaki, Panagiotis Simakis, Carita Paradis, and Andreas Kerren. Identifying the Authors' National Variety of English in Social Media Texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, pages 671–678, Varna, Bulgaria, September 2017. INCOMA Ltd. URL https:\/\/doi.org\/10.26615\/978-954-452-049-6_086.

Alberto Simões, José João Almeida, and Simon D. Byers. Language Identification: a Neural Network Approach. In Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões, editors, *Proceedings of the 3rd Symposium on Languages, Applications and Technologies (SLATE 2014)*, pages 251–265, Bragança, Portugal, 2014.

Gary F. Simons and Charles D. Fennig, editors. *Ethnologue: Languages of the World, Twentieth Edition.* SIL International, Dallas, USA, 2017. Online version: http://www.ethnologue.com.

Anil Kumar Singh. Study of Some Distance Measures for Language and Encoding Identification. In *Proceedings of the Workshop on Linguistic Distances*, pages 63–72, Sydney, Australia, July 2006.

Anil Kumar Singh. *Modeling and Application of Linguistic Similarity.* PhD thesis, International Institute of Information Technology, Hyderabad, 2010.

Anil Kumar Singh and Jagadeesh Gorla. Identification of Languages and Encodings in a Multilingual Document. In *Proceedings of the 3rd ACL SIGWAC Workshop on Web As Corpus (WAC3-2007)*, pages 95–108, Louvain-la-Neuve, Belgium, 2007.

Anil Kumar Singh and Pratya Goyal. A Language Identification Method Applied to Twitter Data. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 26–29, Girona, Spain, 2014.

Navneet Sinha and Gowri Srinivasa. Hindi-English Language Identification, Named Entity Recognition and Back Transliteration: Shared Task System Description. In *Working Notes os Shared Task on Transliterated Search at Forum for Information Retrieval Evaluation FIRE '14*, Bangalore, India, 2014.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale 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*, pages 62–72, Doha, Qatar, October 2014. URL http://www.aclweb.org/anthology/W14-3907.

Clive Souter, Gavin Churcher, Judith Hayes, John Hughes, and Stephen Johnson. Natural Language Identification using Corpus-Based Models. *Hermes, Journal of Linguistics*, 13: 183–203, 1994.

Aleksander Stensby, B. John Oommen, and Ole-Christoffer Granmo. Language Detection and Tracking in Multilingual Documents Using Weak Estimators. In *Proceedings of the Joint IAPR International Workshop Structural, Syntactic, and Statistical Pattern Recognition (SSPR&SPR 2010)*, pages 600–609, Cesme, Izmir, Turkey, 2010.

Erik Sterneberg. Language Identification of Person Names Using Cascaded SVMs. Bachelor's Thesis, Uppsala University, Uppsala, 2012.

Marija Stupar, Tereza Jurić, and Nikola Ljubešić. Language Identification of Web Data for Building Linguistic Corpora. In *Proceedings of the 3rd International Conference on The Future of Information Sciences (INFuture 2011)*, pages 365–372, Zagreb, Croatia, 2011.

Izumi Suzuki, Yoshiki Mikami, Ario Ohsato, and Yoshihide Chubachi. A Language and Character Set Determination Method Based on n-gram Statistics. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(3):269–278, September 2002.

Hidayet Takçı and Ekin Ekinci. Minimal Feature Set in Language Identification and Finding Suitable Classification Method with it. *Procedia Technology*, 1:444–448, January 2012.

Hidayet Takçı and Tunga Güngör. A High Performance Centroid-based Classification Approach for Language Identification. *Pattern Recognition Letters*, 33(16):2077–2084, December 2012.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, Reykjavik, Iceland, 2014.

William John Teahan. Text Classification and Segmentation Using Minimum Cross-Entropy. In *Proceedings of the 6th International Conference Recherche d'Information Assistee par Ordinateur (RIAO'00)*, pages 943–961, Paris, France, 2000.

Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, and Daniela Moctezuma. Gender and Language-variety Identification with MicroTC – Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of CLEF 2017 Evaluation Labs and Workshop*, Dublin, Ireland, September 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1866/`.

Samuel Thomas and Ashish Verma. Language Identification of Person Names using CF-IOF based Weighing Function. In *8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, pages 1769–1772, Antwerp, Belgium, 2007.

Jilei Tian and Janne Suontausta. Scalable Neural Network Based Language Identification from Written Text. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 48–51, Hong Kong, 2003.

Jilei Tian, Juha Häkkinen, Søren Riis, and Kåre Jean Jensen. On Text-based Language Identification for Multilingual Speech Recognition Systems. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, Denver, Colorado, USA, 2002.

Jörg Tiedemann and Nikola Ljubešić. Efficient Discrimination Between Closely Related Languages. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2619–2634, Mumbai, India, 2012.

Christoph Tillmann, Yaser Al-Onaizan, and Saab Mansour. Improved Sentence-Level Arabic Dialect Classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 110–119, Dublin, Ireland, 2014.

Andrija Tomović and Predrag Janičić. A Variant of N-Gram Based Language Classification. In *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence Artificial Intelligence and Human-Oriented Computing (AI* IA 2007)*, pages 410–421, Rome, Italy, 2007.

Dat Tran and Dharmendra Sharma. Markov Models for Written Language Identification. In *Proceedings of the 12th International Conference on Neural Information Processing*, pages 67–70, Taipei, Taiwan, 2005.

Giang Binh Tran, Dat Ba Nguyen, and Bin Thanh Kieu. n-gram based approach for multilingual language identification. Poster. available at `http://comp.mq.edu.au/programming/task_description/VILangTek.pdf`, 2010.

Stephen Tratz, Douglas Briesch, Jamal Laoudi, and Clare Voss. Tweet Conversation Annotation Tool with a Focus on an Arabic Dialect, Moroccan Darija. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 135–139, Sofia, Bulgaria, 2013.

Stephen C. Tratz. Accurate Arabic Script Language/Dialect Classification. Technical report, Army Research Laboratory, 2014.

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska de Jong, and Theo Meder. An Exploration of Language Identification Techniques for the Dutch Folktale Database. In *Proceedings of the LREC workshop Adaptation of Language Resources and Tools for Processing Cultural Heritage*, pages 47–51, Istanbul, Turkey, 2012.

Erik Tromp and Mykola Pechenizkiy. Graph-Based N-gram Language Identification on Short Texts. In *Proceedings of the 20th Annual Belgian Dutch Conference on Machine Learning (Benelearn 2011)*, pages 27–34, The Hague, Netherlands, 2011.

Yoshido Ueda and Seiichi Nakagawa. Prediction for Phoneme/Syllable/Word-Category and Identification of Language using HMM. In *Proceedings of the 1990 International Conference on Spoken Language Processing, volume 2*, volume 2, pages 1209–1212, Kobe, Japan, 1990.

Edvin Ullman. Shibboleth - A Multilingual Language Identifier. Master's thesis, Uppsala University, Uppsala, 2014.

Chris van der Lee and Antal van den Bosch. Exploring Lexical and Syntactic Features for Language Variety Identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 190–199, Valencia, Spain, 2017.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. Language Identification of Short Text Segments with N-gram Models. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3423–3430, Valletta, Malta, 2010.

Vinsensius Berlian Vega and Stéphane Bressan. Continuous-Learning Weighted-Trigram Approach for Indonesian Language Distinction: A Preliminary Study. In *Proceedings of 19th International Conference on Computer Processing of Oriental Languages (ICCPOL 2001)*, Seoul, Korea, 2001a. Oriental Languages Computer Society.

Vinsensius Berlian Vega and Stéphane Bressan. Indexing the Indonesian Web: Language Identification and Miscellaneous Issues. In *Poster Proceedings of the Tenth International World Wide Web Conference (WWW 10)*, Hong Kong, 2001b.

J. Vinosh Babu and S. Baskaran. Automatic Language Identification Using Multivariate Analysis. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, pages 789–792, Mexico City, Mexico, 2005.

Tony Vitale. An Algorithm for High Accuracy Name Pronunciation by Parametric Speech Synthesizer. *Computational Linguistics*, 17(3):257–276, 1991.

John Vogel and David Tresner-Kirsch. Robust Language Identification in Short, Noisy Texts: Improvements to LIGA. In Martin Atzmueller and Hotho Andreas, editors, *Proceedings of the 3rd International Workshop on Mining Ubiquitous and Social Environments (MUSE)*, pages 43–50, Bristol, UK, 2012.

Clare Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. Finding Romanized Arabic Dialect in Code-Mixed Tweets. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, ELRA*, pages 188–199, Reykjavik, Iceland, 2014.

Ada Wan. Leveraging Data-Driven Methods in Word-Level Language Identification for a Multilingual Alpine Heritage Corpus. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 45–54, San Diego, California, 2016.

Martin Wechsler, Sheridan Páraic, and Peter Schäuble. Multi-language Text Indexing for Internet Retrieval. In *Proceeding RIAO'97 Computer-Assisted Information Searching on Internet*, pages 217–232, Montreal, Canada, 1997.

Jennifer Williams and Charlie K. Dagli. Twitter Language Identification Of Similar Languages And Dialects Without Ground Truth. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 73–83, Valencia, Spain, 2017.

Gergely Windisch and László Csink. Language Identification Using Global Statistics of Natural Languages. In *Proceedings of the 2nd Romanian-Hungarian Joint Symposium on Applied Computational Intelligence (SACI)*, pages 243–255, Timisoara, Romania, 2005.

Samantha Wray. Classification of Closely Related Sub-dialects of Arabic Using Support-Vector Machines. In *Proceedings of Language Resources and Evaluation (LREC)*, Miyazaki, Japan, 2018.

Alexandros Xafopoulos, Constantine Kotropoulos, George Almpanidis, and Ioannis Pitas. Language Identification in Web Documents Using Discrete HMMs. *Pattern Recognition*, 37(3):583–594, 2004.

Fei Xia, William D. Lewis, and Hoifung Poon. Language ID in the Context of Harvesting Language Data off the Web. In *+EACL2009*, pages 870–878, Athens, Greece, 2009.

Fei Xia, Carrie Lewis, and William D. Lewis. The Problems of Language Identification within Hugely Multilingual Data Sets. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2790–2797, Valletta, Malta, 2010.

Meng Xuan Xia. Codeswitching Language Identification Using Subword Information Enriched Word Vectors. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 132–136, Austin, TX, USA, 2016. Association for Computational Linguistics.

Fan Xu, Mingwen Wang, and Maoxi Li. Sentence-level Dialects Identification in the Greater China Region. *International Journal on Natural Language Computing (IJNLC)*, 5(6), December 2016.

Mohammad Ali Yaghan. "Arabizi": A Contemporary Style of Arabic Slang. *Design Issues*, 24(2):39–52, 2008.

Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. Text Segmentation by Language Using Minimum Description Length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 969–978, Jeju Island, Korea, July 2012.

Xi Yang and Wenxin Liang. An N-Gram-and-Wikipedia joint approach to Natural Language Identification. In *Proceedings of the 4th International Universal Communication Symposium (IUCS 2010)*, pages 332–339, Beijing, China, 2010.

Noson S. Yanofsky. Towards a Definition of an Algorithm. *Journal of Logic and Computation*, 21(2):253–286, 2011.

Yin-Lai Yeong and Tien-Ping Tan. Language Identification of Code Switching Malay-English Words Using Syllable Structure Information. In *Proceedings of the 2nd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU'10)*, pages 142–145, Penang, Malaysia, 2010.

Yin-Lai Yeong and Tien-Ping Tan. Applying Grapheme, Word, and Syllable Information for Language Identification in Code Switching Sentences. In *Proceedings of the International Conference on Asian Language Processing (IALP 2011)*, pages 111–114, Penang, Malaysia, 2011. IEEE.

Jia-Li You, Yi-Ning Chen, Min Chu, Frank K. Soong, and Jin-Lin Wang. Identifying Language Origin of Named Entity with Multiple Information Sources. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1077–1086, 2008.

Omar F. Zaidan and Chris Callison-Burch. The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41, Portland, Oregon, USA, June 2011.

Omar F. Zaidan and Chris Callison-Burch. Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202, 2014.

Juglar Díaz Zamora, Adrian Fonseca Bruzón, and Reynier Ortega Bueno. Tweets Language Identification Using Feature Weighting. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, pages 30–34, Girona, Spain, 2014.

Marcos Zampieri. Using Bag-of-words to Distinguish Similar Languages: How Efficient Are They? In *Proceedings of the 2013 IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 37–41, Budapest, Hungary, 2013.

Marcos Zampieri. Automatic Language Identification. In *Working with Text: Tools, Techniques and Approaches for Text Mining*, chapter 8, pages 189–205. Elsevier, 2016.

Marcos Zampieri and Binyam Gebrekidan Gebre. Automatic Identification of Language Varieties: The Case of Portuguese. In *Proceedings of The 11th Conference on Natural Language Processing (KONVENS 2012)*, pages 233–237, Vienna, Austria, 2012.

Marcos Zampieri and Binyam Gebrekidan Gebre. VarClass: An Open-source Language Identification Tool for Language Varieties. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3305–3308, Reykjavik, Iceland, 2014.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. N-gram Language Models and POS Distribution for the Identification of Spanish Varieties. In *Proceedings of la 20ème conférence du Traitement Automatique du Langage Naturel (TALN)*, pages 580–587, Sables d'Olonne, France, 2013.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 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, 2014.

Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. Comparing Approaches to the Identification of Similar Languages. In *Proceedings of the*

Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), pages 66–72, Hissar, Bulgaria, 2015a.

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

Marcos Zampieri, Shervin Malmasi, Octavia-Maria Sulea, and Liviu P Dinu. A Computational Approach to the Study of Portuguese Newspapers Published in Macau. In *Proceedings of the Workshop on Natural Language Processing meets Journalism (NLPMJ 2016)*, pages 47–51, New York City, NY, USA, 2016.

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

Andjelka Zecevic and Stasa Vujicic-Stankovic. The Mysterious Letter J. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, pages 40–44, Hissar, Bulgaria, 2013.

Wei Zhang, Robert A. J. Clark, Yongyuan Wang, and Wen Li. Unsupervised Language Identification Based on Latent Dirichlet Allocation. *Computer Speech and Language*, 39: 47–66, 2016.

Anna V. Zhdanova. Automatic Identification of European Languages. In *Revised Papers of the Natural Language Processing and Information Systems - 6th International Conference on Applications of Natural Language to Information Systems, (NLDB 2002)*, pages 76–84, Stockholm, Sweden, 2002.

Ayah Zirikly, Bart Desmet, and Mona Diab. The GW/LT3 VarDial 2016 Shared Task System for Dialects and Similar Languages Detection. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 33–41, Osaka, Japan, 2016.

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

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. TweetLID: A Benchmark for Tweet Language Identification. *Language Resources and Evaluation*, 50 (4):729–766, 2016. ISSN 1574-020X. doi: 10.1007/s10579-015-9317-4. URL `http://dx.doi.org/10.1007/s10579-015-9317-4`.