Text-based Language Identification of Multilingual Names

Oluwapelumi Giwa[1] and Marelie H. Davel[1,2]
[1]Multilingual Speech Technologies, North-West University Vanderbijlpark, South Africa.
[2]CAIR, CSIR Meraka, South Africa.
oluwapelumi.giwa@gmail.com, marelie.davel@gmail.com

***Abstract*— Text-based language identification (T-LID) of isolated words has been shown to be useful for various speech processing tasks, including pronunciation modelling and data categorisation. When the words to be categorised are proper names, the task becomes more difficult: not only do proper names often have idiosyncratic spellings, they are also often considered to be multilingual. We, therefore, investigate how an existing T-LID technique can be adapted to perform multilingual word classification. That is, given a proper name, which may be either mono- or multilingual, we aim to determine how accurately we can predict how many possible source languages the word has, and what they are. Using a Joint Sequence Model-based approach to T-LID and the SADE corpus – a newly developed proper names corpus of South African names – we experiment with different approaches to multilingual T-LID. We compare posterior-based and likelihood-based methods and obtain promising results on a challenging task.**

I. Introduction

Names form a large set of words with complex pronunciation patterns. For example, consider the name 'Wale', which can be a proper name in both Yoruba and English. This name will be pronounced completely different in Yoruba as / w { l e /, compared to /w e I l/ in English (using SAMPA [1] notation). Speech recognition and speech synthesis systems dealing with names require a means to produce accurate pronunciations for these words in order to function properly.

From a human perspective, different speakers tend to pronounce unfamiliar names taking factors such as orthographic irregularity and possible loanword status into account. Specifically, people tend to mimic pronunciations based on what they believe the origin of the name to be, or by replacing unfamiliar phoneme(s) with those they believe are the approximate in their mother tongue [2], [3].

In reality, there are numerous cases where documents or terms (either proper names or generic words) belong to more than one language of origin. Multilingual language identification (LID) is not a well-studied task, with existing text-based language identification (T-LID) techniques focussing either on identifying the single language within a portion of running text, or to a lesser extent, the language identity of isolated words. (See Section II.)

In practical terms, our question then is: given a proper name, can we accurately predict which source languages are relevant, where such names may be either mono- or multilingual? We specifically consider four South African languages using a newly developed proper names corpus. We build on previous work described in [4] using Joint Sequence Models (JSMs) for language tagging of isolated words.

The paper is structured as follows: Section II provides background on related T-LID studies. Section III describes the JSM-LID algorithm that we use as our basic technique. In Section IV, we provide an overview of our approach to the task of multilingual LID, extending the basic JSM-LID technique to cater for multilingual names. Section V presents the two different corpora used in subsequent experiments. Section VI describes our analysis and results. Finally, section VII summarises our findings.

II. Background

Most research in T-LID focuses on the classification problem where a document belongs to one language category in a predefined language space. Documents can be either categorised as long- or short-text segment. According to [5], language identification of long text samples is a solved problem, in which approaches ranging from statistical to pattern recognition algorithms have been applied [6], [7], [8]. Recently, there has been renewed interest in language identification with the focus on short text segments [9], [10], [4], [11], [12], [13].

Text segments such as microblogging data, individual words in isolation and query search text are regarded as short. Traditional T-LID algorithms find this task a difficult one as less contextual information is available. Recent work in this area has especially been directed at microblogging domains [10], [14], [15], [11]. Bergsma *et al*. [11] examined LID on Twitter messages specifically for under-resourced languages, and found that systems trained on out-of-domain data obtained from Wikipedia outperformed other off-the-shelf commercial and academic LID software (TextCat, GoogleCLD, Langid.py). They reported improved performance accuracy using compression-based language models of 97.0% (trained on Wikipedia), 97.4% using a maximum entropy classifier (trained solely on Twitter data), 97.9% using compression-based language models (trained on both Wikipedia and Twitter). Authors also made mention of the factors that contribute to higher performance accuracy such as training data, length of the tweet, and previous information across multiple tweets.

As the text becomes shorter, so the task becomes more difficult. Identification of isolated words (without context) has been approached using techniques such as dictionaries,

[1]The 'Speech Assessment Methods Phonetic Alphabet' is a standard computer-readable notation for phoneme descriptions. See [1].

character n-gram language model, JSMs, support vector machines and conditional random fields [16], [4], [17], [18].

In previous work [4], JSM (Dependinga pronunciation modelling technique) was compared against SVM models for T-LID of words in isolation. Experiments conducted on four South African languages (Afrikaans, English, Sesotho and iziZulu) reported competitive results. The JSM-based system obtained an F1-measure of 97.2% as compared to a state-of-the-art SVM technique with an F1-measure of 95.2%. King and Abney [18] used a weakly supervised approach for identifying the languages of single words in a multilingual document. They experimented with different ranges of data sizes and reported that conditional random fields models trained with generalised expectation outperformed sequence classifiers.

From isolated words to proper name identification, task complexity increases. Konstantopoulos [19] examined language identification of proper names. He experimented with soccer player names obtained from 13 languages. He reported an initial average F_1 score of 27% when tested on a general n-gram language model. With a more discriminated training data based on short sizes, an average F_1 score of 50% was obtained on last names and 60% on first names. In related work, Li $et\ al$ [20] used an n-gram language model to identify proper names in English, Chinese and Japanese. They reported an overall accuracy of 94.8% when classifying names amongst these three languages. (As these three languages are not closely related, the classification task becomes easier, explaining the high accuracy achieved.)

To the best of our knowledge, no studies are available that address the task of identifying multiple source languages of names in isolation. The closest related task addresses LID for multilingual documents (where a single document can belong to more than one language class). Approaches to this task include word-level identification [17], [18], vector-space model [21], minimum description length principle [22] and monolingual block segmentation [23], [24]. Within this context, Nguyen and Dogruoz [17] considered word-level classification in order to discriminate between Dutch and Turkish in a multilingual online discussion. They experimented with language model classification, logistic regression classification, dictionaries and conditional random fields; they reported the best performance using language models, while contextual information remains beneficial.

III. Joint Sequence Models for T-LID

In this section, we first describe the JSM algorithm and how it can be applied to T-LID, before extending it to the multilingual case in Sections IV and VI.

Joint Sequence Models were defined by Bisani and Ney in [25]. Initially developed for grapheme-to-phoneme (G2P) modelling, the technique is built around the concept of a 'graphone', an m-to-n alignment between small sections of graphemes and phonemes that form the basic units for probability modelling. Both the possible alignments and the graphones themselves are estimated through embedded maximization using a training dictionary. The probability of one unit occurring given the other(s) are similarly estimated using the same training data. The application of JSMs to LID is described in [4]. Below we provide an overview of the data preparation and training, as well as transcription phases. For more detail see [25] and [4].

A. Training phase

JSMs are typically used for pronunciation prediction. In order to generate a dictionary for T-LID training, all words with their corresponding language identifiers are added to a 'pseudo dictionary'. To simplify the co-segmentation task, a one-to-one letter to LID mapping is adopted whereby each grapheme corresponds to a single language identifier. For example, consider Table I where '#' represents word boundary markers, 'E' represents to English, 'Z' represents isiZulu, 'A' represents Afrikaans, and 'S' represents Sesotho.

TABLE I
EXAMPLE OF A JSM DICTIONARY RECAST FOR THE T-LID TASK

Word	Pronunciation
#school#	E E E E E E E E
#lekker#	A A A A A A A A
#zuma#	Z Z Z Z Z Z
#lebo#	S S S S S S

As noted in [4], parameter choices such as graphone length, discounting, m-gram model order, and how models are initialised, influence performance accuracy of the JSM model. For this work, 1-1 graphones are trained, which means that the minimum and the maximum number of graphemes to LID mappings allowed per graphone length is 1. As models typically saturate before the 8^th order, an 8^th order model is used. Discounting is allowed during training to handle unseen token and avoid overfitting. All held-out data used for parameter estimation are folded back to the training set.

B. Transcription phase

In standard JSMs, a forward algorithm is used to compute the joint probability, $p(g, \varphi)$, of a co-segmentation between a grapheme sequence g and phoneme sequence φ. The probability of a source sequence g is in principle determined by summing all the matching graphone sequences across the sequence path. This value can also be approximated by simply taking the maximum value:

$$p(\varphi \mid g) = \frac{max_{q \in S(g,\varphi)} p(q)}{p(g)} \quad (1)$$

where $S(g, \varphi)$is the set of all co-segmentations of g, and $p(q)$ represents the probability distribution over the sequences of graphones. (All probabilities as estimated during training).

JSMs allow the ability to generate different pronunciation variants. That is, for each word, a number of pronunciation variants may be produced; the path posterior probability given the word is used to select the winning candidate amongst the variants. As in the T-LID case, a variant consists of an LID string, ambiguity is resolved either by selecting the language identifier with the highest frequency counts

(within the variant) or by summing the language-specific log-probabilities internally to the word. (In Section VI-B, an alternative is proposed.)

IV. Approach

We approach the multilingual LID task by first selecting a technique that has been applied successfully to monolingual words in the past. Specifically, we select JSM-based LID, which demonstrates competitive performance for isolated word LID [4].

Two main issues must be addressed:

1) *The data to train the JSM models.* Is it better to use matching data (names) even if this data set is very small, or better to use a significantly larger set, even if it is unmatched (generic words)?
2) *Options for extending the technique for multilingual data.* JSMs can generate variant outcomes, as well as the likelihood and posterior probability given the word, per outcome. How well do these values predict true multilingual words?

We obtain empirical results by experimenting with different data sets and thresholds. We select the SADE multilingual name corpus [26] and the NCHLT *in-lang* dictionaries [27] for experimentation. These data sources are described in more detail in Section V. We discuss and experiment with different classification options in Section VI. We first propose an improvement to the monolingual classification technique, before experimenting with different approaches to multilingual classification.

As performance measures we use the standard definitions of precision, recall and F1-measure, that is:

$$\text{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

with TP the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. Note that 'recall' is the same as True Positive Rate (TPR). When a monolingual test set is used, these three measures are equivalent and we refer simply to accuracy.

When comparing threshold-based techniques we also evaluate the Receiver Operating Characteristic (ROC) by plotting the TPR (eq. 3) against the True Negative Rate (TNR) rather than against the False Positive Rate, as also often done, with

$$TNR = \frac{TN}{TN + FP} \tag{5}$$

In a ROC curve, the further the curve from the diagonal to the upper right hand side the better the performance of the system. Data points located far from the center of the origin depicts poor performance.

V. Data

We use two data sets in this analysis: the South African Directory Enquiries (SADE) corpus [26] to obtain tagged samples of multilingual names, and the NCHLT dictionaries [27] to obtain samples of generic monolingual words in matching languages. We analyse LID performance for four South African languages, namely Afrikaans, English, isiZulu and Sesotho.

The SADE corpus was developed to improve directory enquiry applications in South Africa. The corpus contains audio samples from multilingual speakers producing different proper names, and reflects a range of scenarios directory enquiries system could encounter. SADE data prompts were developed using publicly available names from a combination of Internet queries, personal names from a local tertiary institution (North-West University) and volunteers. Each name is annotated with a number of tags, including the most probable source language. We use only the word lists, and the LID tags of the v1.1. corpus in the current analysis.

The NCHLT dictionaries were developed in parallel with the NCHLT speech corpus [28]. For this work, we only use the word lists from the dictionaires, consisting of 15 000 unique words per language. These were estimated to be frequent words based on available corpus counts, and LID accuracy was verified using automated spell checkers and language practitioners [27]. We use the same edited lists as used in [4], where a second round of higher precision, lower recall spell-checking was performed for three of the languages (Afrikaans, isiZulu and English). These lists do not contain multilingual words.

VI. Analysis and Results

We first analyse the distribution of multilingual words in the SADE corpus and create a suitable train and test set (Section VI-A) before applying JSMs to the monolingual test case (Section VI-B). The LID technique is then extended to cater for multilingual words in Section VI-C.

A. Data analysis

While the SADE corpus contains words in a large variety of languages, we only use the subset consisting of words tagged with the four target languages (Afrikaans, English, Sesotho and isiZulu). While the actual data set is almost exclusively bilingual by which Afrikaans - English definitely dominate. Per language, the number of unique words, average word length and total character count are displayed in Table II. The names contained in the corpus are not only person names, but also include the names of songs, restaurants and places, for example. These often consists of phrases (such as '*The Hillside Tavern*'). The resulting set of words is therefore a mixture of names and some generic words.

While the majority of the words are monolingual, a significant percentage of them (9.3%) were identified as multilingual. The distribution between mono- and multilingual words per language is shown in Table III. As before,

TABLE II

SADE CORPUS: LANGUAGE DISTRIBUTION AND WORD STATISTICS.

Language	Word Count	Average word length	Character count
Afrikaans	1 050	6.9	7 308
English	6 634	7.4	48 733
Sesotho	465	7.8	3 612
isiZulu	458	8.1	3 689

the number of unique words, average word length and total character count are displayed.

TABLE III

SADE CORPUS: MONO- AND MULTI-LINGUAL DISTRIBUTION AND WORD STATISTICS.

	Word Count		Average word length		Character count	
Language	Mono	Multi	Mono	Multi	Mono	Multi
Afrikaans	449	601	8.5	5.8	3 808	3 500
English	5 980	654	7.5	5.7	4 974	3 759
Sesotho	411	54	8.1	4.9	3 345	267
isiZulu	401	57	8.4	5.3	3 387	302

Due to the small word count of our isiZulu language, we randomly select a monolingual subset, per language, that equals to 401 unique words. For verification purpose, we ask language practitioners to review language tags of our newly extracted subset together with all the multilingual word-list shown in Table III. We analyse the results obtained from the language practitioners and observed few changes across words and their corresponding language IDs. In order to create a balance training set using only monolingual words, we randomly select a subset of 321 per language. Without restricting the allowed range of words per language in the test set, all remaining selection is random. The resulting training and test set partition is displayed in Table IV.

Most of the multilingual words in the corpus are bilingual, with a very small set of 3-lingual and a single 4-lingual word (‘sale’). Examples of 3-lingual words include ‘tutu’, ‘tone’ and ‘pole’. The exact distribution of words are shown in Fig. 1. Note that the figure is restricted to 800 words, even though the first bar exceeds this number (with a value of 7241 words).

The largest confusability exists between English and Afrikaans words, as can be seen from Table V, which lists the number of bilingual words in the SADE test set. For example, the word count of 314 represents the total number of words that do exist in Afrikaans and English only, 15 represents total word count that exist both in Sesotho and isiZulu only.

TABLE IV

NUMBER OF MONO- AND MULTILINGUAL WORDS IN THE SADE TRAIN AND TEST PARTITIONS.

	Training set		Test set	
Language	Mono	Multi	Mono	Multi
Afrikaans	321	-	155	329
English	321	-	137	348
Sesotho	321	-	98	34
isiZulu	321	-	99	34

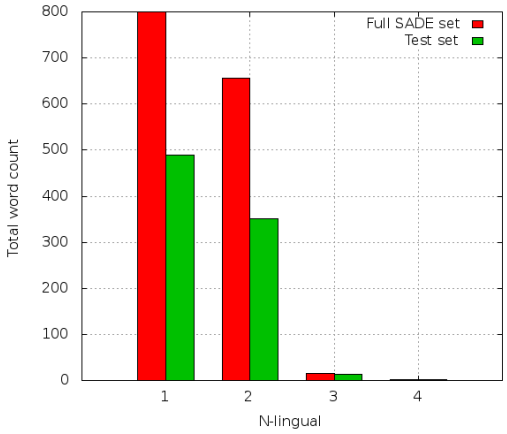

Fig. 1. Number of 1-, 2-, 3- and 4-lingual words in the SADE full and test data sets.

TABLE V

LANGUAGE IDENTITIES OF BILINGUAL WORDS IN THE SADE TEST SET

Languages		Word count
Afrikaans	English	314
Sesotho	isiZulu	15
English	isiZulu	13
English	Sesotho	7
Afrikaans	Sesotho	2

B. LID of monolingual names

For the generic model, we use a 40K-word subset of the NCHLT data. Our objective is to make use of an already existing model based on previous work done in [4]. For comparative purposes, the same statistics as shown for the SADE data in Tables II and III are displayed for the NCHLT data in Table VI. This provides us with two different training data sets: a small (1 284) SADE set, and a large (40K) NCHLT set.

TABLE VI

NCHLT 40K SUBSET: LANGUAGE DISTRIBUTION AND WORD STATISTICS.

Language	Word count	Average word length	Character count
Afrikaans	10K	10.5	104 531
English	10K	7.9	79 768
Sesotho	10K	8.3	82 537
isiZulu	10K	9.4	93 617

We first obtain results using the identical technique as described in [4]. Specifically, we sum the language-specific log-probabilities internally to the word. The latter technique (referred to as ‘logprob sum’) produced the most accurate results in [4]. We also report on an extension to this technique, where each variant is forced to be monolingual internally. That is, a 5-letter word will only be tagged as ‘EEEEE’ or ‘ZZZZZ’: a combination such as ‘EEZZZ’ is not allowed. In the 4-language task, this means each word would produce at most 4 variants, each with an associated posterior probability. These posteriors are then used to select the winning candidate, and any mixed variants simply discarded. This technique is referred to as ‘forced pron’ below.

TABLE VII

LID RESULTS FOR THE SADE MONOLINGUAL TEST SET, USING DIFFERENT TRAINING DATA SETS AND JSM-BASED TECHNIQUES.

Data set	Technique	Accuracy
NCHLT 40K	logprob sum	77.96
NCHLT 40K	forced pron	78.16
SADE	logprob sum	79.39
SADE	forced pron	81.02

In Table VII we report on results. Interestingly, the much smaller SADE training data clearly fits the test data better and produces more accurate results. The new 'forced pron' technique also shows a small but consistent improvement, across all measures.

C. LID of multilingual names

Using the SADE models and the 'forced pron' LID technique, we evaluate two approaches for determining when a word may be truly multilingual:

- We define an absolute threshold based on the posterior probability: any variants with a posterior probability higher than the threshold are accepted as additional source languages. We refer to this technique as 'absolute posterior' further.
- We define a relative threshold based on the log likelihood: any variants with a relative likelihood within the range of the best variant are accepted as additional source languages. We refer to this technique as 'relative likelihood' further.

In both cases, the best performing variant is automatically selected: thresholds are only used to determine whether more than one source language may potentially apply.

The ROC curves for both these methods are displayed in Fig. 2. (The closer the graph to the top right corner, the more accurate the technique.) The full test set from Table IV is used. As expected, the two techniques provide very similar results, with optimal F_1 scores of 79.99% and 79.85% obtained, by 'absolute posterior' and 'relative likelihood', respectively.

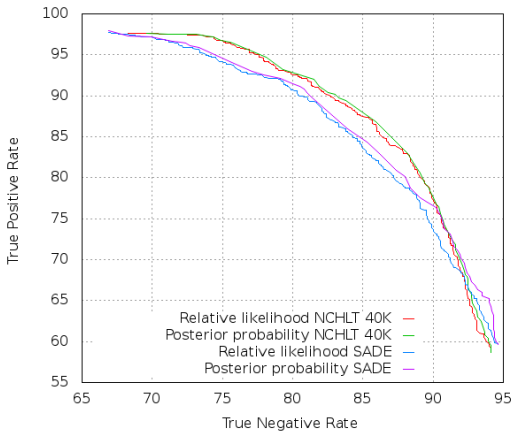

Fig. 2. ROC curves for the SADE combined test set comparing the 'absolute posterior' and 'relative likelihood' approaches.

From the ROC curve in Fig. 2 it is clear that the techniques perform as expected, but how well do they perform? In order to answer this question, we consider a simple baseline whereby we select first the single best variant ('top-1'), and then the two best variants ('top-2'). In the first case, all words are therefore treated as if they are monolingual, and all additional source languages will automatically be 'false rejects'. In the second case, a large number of superfluous source languages will be hypothesized, and many 'false accepts' are accepted. We compare the results from these four methods in Table VIII.

TABLE VIII

COMPARING DIFFERENT MULTILINGUAL CLASSIFICATION APPROACHES USING THE SADE COMBINED TEST SET.

Approach	Model	Recall	Precision	F-measure
top-1	NCHLT 40K	58.67	84.78	69.35
top-1	SADE	59.56	86.07	70.40
top-2	NCHLT 40K	91.89	66.51	77.17
top-2	SADE	91.98	66.84	77.42
relative likelihood	SADE	84.60	75.60	79.85
absolute posterior	SADE	84.80	75.70	79.99
relative likelihood	NCHLT 40K	86.00	77.60	81.58
absolute posterior	NCHLT 40K	86.20	77.80	81.79

From Table VIII it is clear that this is a difficult task: the first baseline result ('top-1') produces an F-measure of only 69.35% and 70.40% using the NCHLT 40K and SADE models, respectively. The second baseline approach ('top-2') shows improvement over the 'top-1' technique, where a recall value of 92% (obtained on both NCHLT 40K and SADE models) means a large number of the correct tags were correctly identified. However, this high recall value comes at a cost: specifically a precision that falls to 67%. As we accommodate more tags based on the number of variants, there is a high likelihood of also identifying the wrong labels. In contrast to this, the 'top-1' approach achieves better precision value, but with lower recall values.

We improve on the baseline results, where we leverage the trade-off between the 'false rejects' and 'false accepts' related to recall and precision respectively. For both approaches, we select an optimum point across different threshold values where TNR equals TPR, and compute the recall and precision values. The optimum threshold value where TNR equals TPR for 'relative likelihood' is -3.17, while the 'absolute posterior' is 0.026. Converting the optimum threshold value of 'relative likelihood' to a probability form, produces a value of 0.042, which is (as expected) close to the optimal value returned by the 'absolute posterior' approach.

Interestingly, we observe that using the NCHLT 40K model on 'absolute posterior' produces the best performance for multilingual classification of proper names. However, this contradicts our initial observation concerning monolingual classification, where the SADE model produced the best performance. This result shows that we cannot judge the performance of a multilingual classifier by only considering its monolingual classification accuracy.

VII. Conclusion

This work focussed on the multilingual classification of proper names, a task that has not been well-studied to date. Although our work is targeted at four South African languages, the techniques utilised are language independent.

First, we proposed an improvement to the monolingual classification of words using JSMs, building on earlier work described in [4]. By forcing JSMs to produce output strings that are associated with a single language, we obtain more trustworthy posteriors to analyse. We compared this approach to the best available to date ('logprob sum') and observed an improvement in F-measure from 79.39% to 81.02%, training and testing on the same set of proper names.

In order to classify proper names as multilingual, we experiment with two baseline methods ('top-1' and 'top-2') where both approaches produce a tradeoff between recall and precision. To strike a balance between our two metric values, we proposed two new techniques ('relative likelihood' and 'absolute posterior'). While the difference between the two new methods is statistically insignificant, both outperform the baseline with 'absolute posterior' using the NCHLT models producing an F-measure of 81.79%.

Finally, we observe that the identification performance on monolingual proper names does not necessarily translate to similar performance for multilingual classification. For the LID of monolingual names, models trained on SADE outperformed models trained on NCHLT 40K with an F-measure of 81.02% (compared to 78.16%). For LID of multilingual names, the NCHLT 40k models produced better results than SADE. (81.79% compared to 79.99%).

In conclusion, we have shown that even though LID of multilingual proper names is a challenging task, an adapted version of JSMs provide good classification accuracy.

VIII. Acknowledgment

This work was partially supported by the National Research Foundation. Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the NRF do not accept any liability in regard thereto.

References

[1] D. Gibbon, R. Moore, and R. Winski, *Handbook of standards and resources for spoken language systems*. Walter de Gruyter, 1997.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10–11, pp. 763–786, 2007.

[3] F. Stouten and J.-P. Martens, "Dealing with cross-lingual aspects in spoken name recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding ASRU*, 2007, pp. 419–424.

[4] O. Giwa and M. H. Davel, "Language identification of individual words with joint sequence models," in *Proc. 15th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 14-18 September, Singapore, 2014, pp. 1400–1404.

[5] P. McNamee, "Language identification: a solved problem suitable for undergraduate instruction," *Journal of Computing Sciences in Colleges*, vol. 20, no. 3, pp. 94–101, 2005.

[6] Y. Chen, J. You, M. Chu, Y. Zhao, and J. Wang, "Identifying language origin of person names with n-grams of different units," in *Proc. International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2006, pp. 729–732.

[7] C. Kruengkrai, P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara, "Language identification based on string kernels," in *Proc. ISCIT*, 2005, pp. 926–929.

[8] M. Padró and L. Padró, "Comparing methods for language identification," *Procesamiento del lenguaje natural*, vol. 33, pp. 155–162, 2004.

[9] T. Vatanen, J. J. Väyrynen, and S. Virpioja, "Language identification of short text segments with N-gram models." in *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, 2010, pp. 3423–3430.

[10] E. Tromp and M. Pechenizkiy, "Graph-based n-gram language identification on short texts," in *Proc. of the 20th Machine Learning conference of Belgium and The Netherlands*, The Hague, Netherlands, 2011, pp. 27–34.

[11] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson, "Language identification for creating language-specific twitter collections," in *Proc. of the Second Workshop on Language in Social Media*, 2012, pp. 65–74.

[12] M. Lui, J. H. Lau, and T. Baldwin, "Automatic detection and language identification of multilingual documents," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 27–40, 2014.

[13] G. R. Botha and E. Barnard, "Factors that affect the accuracy of text-based language identification," *Computer Speech & Language*, vol. 26, no. 5, pp. 307–320, 2012.

[14] S. Carter, W. Weerkamp, and M. Tsagkias, "Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text," *Language Resources and Evaluation*, vol. 47, pp. 195–215, 2013.

[15] M. Goldszmidt, M. Najork, and S. Paparizos, "Boot-strapping language identifiers for short colloquial postings," in *Proc. of the Machine Learning and Knowledge Discovery in Databases*, 2013, pp. 95–111.

[16] O. Giwa and M. H. Davel, "N-gram based language identification of individual words," in *Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Johannesburg, South Africa, 2013, pp. 15–21.

[17] D. Nguyen and A. S. Dogruoz, "Word level language identification in online multilingual communication," in *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA, 2014, pp. 857–862.

[18] B. King and S. P. Abney, "Labeling the languages of words in mixed-language documents using weakly supervised methods." in *Proc. of the NAACL-HLT*, 2013, pp. 1110–1119.

[19] K. Stasinos, "What's in a name? quite a lot," in *Proc. of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP-07)*, Borovets, Bulgaria, 2007.

[20] H. Li, K. C. Sim, J.-S. Kuo, and M. Dong, "Semantic transliteration of personal names," in *Proc. of the Annual Conference of the Association for Computational Linguistics*, 2007, pp. 120–127.

[21] J. M. Prager, "Linguini: Language identification for multilingual documents," in *Proc. of the 32nd Annual Hawaii International Conference on Systems Sciences. HICSS-32*, 1999, pp. 11–16.

[22] H. Yamaguchi and K. Tanaka-Ishii, "Text segmentation by language using minimum description length," in *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju Island, Korea, 2012, pp. 969–978.

[23] W. J. Teahan, "Text classification and segmentation using minimum cross-entropy," in *Proc. of the 6th International Conference Recherche d'Information Assistee par Ordinateur (RIAO' 00)*, 2000, pp. 943–961.

[24] T. Mandl, M. Shramko, O. Tartakovski, and C. Womser-Hacker, "Language identification in multi-lingual web-documents," in *Proc. of the 11th International Conference on Applications of Natural Language to Information Systems (NLDB 2006)*, Klagenfurt, Austria, 2006, pp. 153–163.

[25] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[26] Thirion, Jan W.F. and van Heerden, Charl and Giwa, Oluwapelumi and Davel, Marelie H., "The South African Directory Enquiries (SADE) corpus," in preparation.

[27] M. H. Davel, W. D. Basson, C. van Heerden, and E. Barnard, "NCHLT Dictionaries: Project Report," Multilingual Speech Technologies, North-West University, Tech. Rep., May 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/home

[28] E. Barnard, M. H. Davel, C. J. V. Heerden, F. D. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, 2014, pp. 194–200.