Arabic Dialect Identification in the Context of Bivalency and Code-Switching

Mahmoud El-Haj[1], **Paul Rayson**[1] **and Mariam Aboelezz**[2]
[1]School of Computing and Communications, Lancaster University, UK and [2]British Library, UK
{m.el-haj, p.rayson}@lancaster.ac.uk, mariam.aboelezz@bl.uk

Abstract

In this paper we use a novel approach towards Arabic dialect identification using language bivalency and written code-switching. Bivalency between languages or dialects is where a word or element is treated by language users as having a fundamentally similar semantic content in more than one language or dialect. Arabic dialect identification in writing is a difficult task even for humans due to the fact that words are used interchangeably between dialects. The task of automatically identifying dialect is harder and classifiers trained using only n-grams will perform poorly when tested on unseen data. Such approaches require significant amounts of annotated training data which is costly and time consuming to produce. Currently available Arabic dialect datasets do not exceed a few hundred thousand sentences, thus we need to extract features other than word and character n-grams. In our work we present experimental results from automatically identifying dialects from the four main Arabic dialect regions (Egypt, North Africa, Gulf and Levant) in addition to Standard Arabic. We extend previous work by incorporating additional grammatical and stylistic features and define a *subtractive bivalency profiling* approach to address issues of bivalent words across the examined Arabic dialects. The results show that our new methods classification accuracy can reach more than 76% and score well (66%) when tested on completely unseen data.

Keywords: Arabic, bivalency, language identification, dialects, machine learning, NLP

1. Introduction

In natural language processing, the problem of detecting the language of a given text is called language identification or language guessing. In early work, relatively simple approaches employing character n-grams proved to be successful (Cavnar and Trenkle, 1994; Dunning, 1994; Souter et al., 1994). More recently, this has been seen as a classification problem where machine learning is used to distinguish between languages (Gupta et al., 2015). The vast majority of language identification research has focused on differentiating between languages. In this paper, we instead focus on differentiating regional varieties of the same language (i.e. dialects), taking Arabic as our case study.

Automatically identifying dialects could prove fruitful in fields such as natural language processing, corpus linguistics and machine translation (Zaidan and Callison-Burch, 2014). The process of hiring human participants to identify dialects is very costly and it is a time consuming job. Therefore, machine automation could work as a quick and cheap alternative provided we can create an effective mix of new methods with the appropriate dataset.

In this study, we tackle the problem of automatically identifying Arabic dialects using a variety of approaches in order to address bivalency and dialectal written code-switching (Habash et al., 2008; Biadsy et al., 2009) which pose significant challenges for existing approaches. We apply our novel *Subtractive Bivalency Profiling (SBP)* approach to address the issue of bivalent words across the Arabic dialects examined here. The results show that our new methods can achieve good levels of accuracy on unseen data.

Bivalency is defined by Woolard and Genovese (2007) as the "simultaneous membership of a given linguistic segment in more than one linguistic system in a contact setting". It is typically a feature of linguistic codes that are closely related to each other, like Standard Arabic and the various Arabic colloquial varieties. Woolard uses the term 'strategic bivalency' to refer to deliberate linguistic manipulation that makes it nearly impossible to classify a segment of speech as belonging to one code or the other. She originally introduced the term bivalency to talk about spoken language use (namely Spanish and Catalan), but it has since been extended to writing. For example, Mejdell (2011) later studied strategic bivalency in written Arabic with respect to Standard Arabic and Egyptian Arabic.

Bivalency is a hallmark feature of written Arabic, especially in the case of single-word or short utterances such as "القلم علَى المكتب" (the pen is on the table). The three words in this utterance can be found - with the same semantic content - in all major Arabic dialects. What makes bivalency more common in writing than speech is that regional variants such as 'qalam', 'galam' and 'alam' - which are easily distinguished in speech - are all likely to be represented using the Arabic writing system as "قلم". Hence, written bivalency (bivalency hereafter) is not simply the result of overlap in vocabulary but also the loss of important linguistic information when different pronunciations are encoded using the same standard representation in Arabic script.

2. Related Work

For differentiating texts at the language level, comparing the relative ranks of character n-grams has proved to be a very successful approach (Cavnar and Trenkle, 1994). Recent research has tackled language identification in noisy settings such as online forums and social media using ensemble methods (Lui and Baldwin, 2014) or more complex statistical approaches (Abainia et al., 2016). Much other research focuses on language identification or recognition from speech signals but that is out of scope for this paper. In the area of corpus linguistics, language identification is not studied directly, but the field has a long history of comparing language varieties and has developed a number of approaches to explore this issue e.g. keywords used in an American versus British English study (Hofland and Johansson, 1982) and multidimensional approaches (Biber, 1988).

Previous work on the more fine-grained task of dialect identification itself is much more scarce with some reported research on African-American English (Blodgett et al., 2016) and European versus Brazilian Portuguese (Laboreiro et al., 2013).

Zaidan (2014) created an Arabic resource of dialect annotation using Mechanical Turk crowdsourcing. The annotators labeled 100,000 sentences defining the Arabic dialect used in writing. They trained a classifier to identify dialectal Arabic in text harvested from online social media. The dialects used to train their classifier were Egyptian, Gulf, Levant, Iraqi and Maghrebi (also known as North African Arabic). The probabilistic classification models used words and character n-gram features. Considering the overlap between dialects the training accuracy did not perform better than 88%. In our work, we hypothesise that using additional grammatical and stylistic features would outperform relying on n-gram features alone to classify Arabic dialects.

Elfardy (2014) trained a supervised classifier to distinguish between Modern Standard Arabic and Egyptian dialect extracting n-gram and token based features to class each word in a sentence whether it is MSA, Egyptian or out-of-vocabulary (OOV). The best configuration classifier achieved an accuracy of just above 80%. For a binary classification task we expect the classifier to perform better. For example, we ran a simple n-gram binary classifier using just MSA and Egyptian, and the classifier achieved near 99% accuracy on training and above 95% when tested on unseen data.

3. Dataset

The dataset used in our research covers four major Arabic dialect groups: Egyptian (EGY), Levant (LAV), Gulf (GLF), and North African (NOR). The dataset also includes Modern Standard Arabic (MSA)[1]. Apart from NOR, all the other dialects were collected from the Arabic Commentary Dataset (AOC) (Zaidan and Callison-Burch, 2014). We randomly selected commentaries written in MSA, EGY, GLF and LAV. Iraqi and Maghrebi (NOR) sections from that corpus do not provide enough sentences for our experiments.

In their preparation of the AOC dataset, it was annotated through hiring online participants on Mechanical Turk[2]. The participants annotated the dataset through answering two questions a) how much dialect is in the sentence, and b) which Arabic dialect the writer intends. It is worth noting that we only selected sentences where the answers to the first question is 'mostly dialect' and where either EGY, GLF, MSA or LAV is the answer to the second question. We did not include NOR (called Maghrebi in AOC) as there were not enough NOR sentences where the answer to the first question is 'mostly dialect'. Instead we supplemented the collection with NOR dialect from Tunisian Arabic which is a free online corpus of Tunisian (North African) Arabic[3]. We randomly selected sentences from the *Internet Forums* category so that it is consistent with the AOC dataset.

During our initial examination of the corpus data, we noticed bivalency between the dialects in the dataset. We also noticed that many writers used a combination of dialect and MSA. This is common in political debates where readers comment on a political news article and others respond to them.

The dataset is rich with bivalent words such as "ترجمة" [translation] and "النّاس" [people]. There is also frequent code-switching to MSA, with phrases such as "حتّى هذه اللّحظة" [Until this moment] and "لم يسبق ان" [never before] which are rarely used in dialect conversations. We later describe how this played a vital role in our dialect identification process. Table 1 shows the count of sentences (instances/samples) and words for each class.

Dialect Label	Sentences	Words
GLF	2,546	65,752
LAV	2,463	67,976
MSA	3,731	49,985
NOR	3,693	53,204
EGY	4,061	118,152
Total	16,494	355,069

Table 1: Training data size

4. Automatic Dialect Identification (ADID)

For the purpose of this task we trained different text classifiers using four algorithms: Naïve Bayes, Support Vector Machine (SVM), k–Nearest Neighbor (KNN) and Decision Trees (J48).

4.1. Baselines

For the first baseline, we created a simple classifier that always selects the most frequent class (EGY in this case). As a more intelligent baseline, we extracted simple word-level n-gram features selecting unigram, bigram and trigram contiguous words using a Naïve Bayes classifier. The second baseline classifiers' accuracy was expected to be an improvement over the most frequent class approach.

4.2. Feature Extraction

To help the classifier distinguish between the dialects more accurately, we extracted more linguistically informed features in addition to our subtractive bivalency profiling method. The selected features fall into two groups: grammatical and stylistic.

4.2.1. Grammatical Features

In order to extract grammatical knowledge from the training data we used the Stanford Part of Speech (POS) Tagger[4] to annotate the text with part-of-speech tags. The POS tagger was trained on an MSA dataset but we judged it to be appropriate enough for our experiments. Key differences

[1]The Arabic Dialects Dataset is freely available for research purposes and can be directly downloaded from http://www.lancaster.ac.uk/staff/elhaj/corpora.htm

[2]http://www.mturk.com

[3]http://www.tunisiya.org/

[4]http://nlp.stanford.edu/software/tagger.shtml

will be in the sentence structure and the introduction of dialect words that may not have appeared in the MSA training data. We expect this may make the tagger more error-prone, but we wish to understand whether it can still help in distinguishing between dialects.

Using the annotated training data introduced in section 3., we extracted a number of grammatical features. *Tag frequency* refers to the frequency of each tag found in the POS tagset while *uniqueness* refers to the number of tag types introduced in the text. In addition to the tag frequencies, we also extracted features which are counts of function words of the following types: adverbs, adverbials, conjunctions, demonstratives, modals, negations, particles, prepositional, prepositions, pronouns, quantifiers, interrogatives and comparatives. Each list contains function words/tags related to that category (Garcia-Barrero et al., 2013; Ryding, 2014).

4.2.2. Stylistic Features

In addition to grammatical features we also extracted two stylistic features, namely a readability metric and Type-Token-Ratio (TTR) which are used elsewhere in authorship identification (Holmes, 1994). TTR is the ratio obtained by dividing the total number of different words (i.e. types) occurring in a text by the total number of words (tokens). Higher TTR indicates a high degree of lexical variation. We calculated TTR by simply dividing the number of types by the number of tokens in each instance (Holmes, 1994):

$$TTR = \frac{types}{tokens}$$

We normalised the output by dividing by the number of sentences in each instance, this was achieved by using the Stanford Arabic sentence splitter[5].

Furthermore, we measured the readability of the text using the OSMAN readability metric (El-Haj and Rayson, 2016). In addition to providing a readability score between 0 (hard to read) and 100 (easy to read), OSMAN also provides information about the number of syllables, hard words (words with more than 5 letters), complex words (>4 syllables) and Faseeh (aspects of script usually dropped in informal Arabic writing).

4.2.3. Subtractive Bivalency Profiling

As mentioned earlier the dataset contains a high level of language bivalency, which is typical when speakers switch between closely related language varieties. We used an approach influenced by earlier work in corpus linguistics in order to select features to study the closeness and homogeneity between the texts in the different classes. We have therefore devised a new method which we have dubbed *Subtractive Bivalency Profiling* (SBP). When examining the frequency lists for each dialect, we noticed that writers occasionally switch to MSA in an apparent bid to invoke formality and/or authority. In order to use bivalency and written code-switching as features in the classification process, we created dialect-specific frequency lists to distinguish the vocabularies spoken in each dialect compared to MSA. The frequency lists we created are of two types: a) dialect SBP list, and b) MSA written code-switching list. In the former list we worked on identifying and removing bivalent words between dialects aside from MSA leaving us with more fine-grained dialectal lists. We then found bivalent words between dialects and MSA, which we refer to as MSA written code-switching. Dialect SBP lists were created using an independent dialectal dataset that has not been exposed to the training process. The independent dataset called DART[6] contains more than 24,000 Arabic sentences labeled into 5 Arabic dialects (Egyptian, Gulf, Levantine, Iraqi and Maghrebi (North African)) matching the dialects we are working with for this paper (we have taken Iraqi out). We created a dialect SBP list for each dialect (in addition to MSA). This was done by creating a unique (with no duplicates) frequency list for each dialect in DARTS removing from that list any bivalent words between that dialect and any of the remaining dialects.

The MSA written code-switching list was created using an independent list of MSA sentences from the United Nations (UN) Corpus[7]. We created a frequency list for the UN corpus and detected bivalency with each of the other four dialects. Here we keep the bivalent words between MSA and each of the other dialects creating a MSA written code-switching list for each dialect.

5. Feature Selection

The count of the selected features including each entry of the frequency and grammatical lists was 50 divided into 3 categories as in Subsection 4.2.. Training the algorithms using this number of features took a significant amount of time thereby making it difficult to attempt many different algorithms. In order to simplify the model, shorten the training time and enhance generalisation to reduce overfitting, we reduced the number of features using machine learning feature selection technique as explained below.

5.1. Classifier Subset Evaluator

To reduce the number of features we used WEKA[8] Classifier Subset Evaluator which evaluates attribute subsets on training data and uses a SVM classifier to estimate the merit of a set of attributes. This helps evaluate the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. This was combined with a Best First Search (BFS) which searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. The attribute selection process selected 8 features as the most predictive ones. The selected features are: SBP (MSA_SBP, EGY_SBP, GLF_SBP, LAV_SBP, NOR_SBP), Grammatical (Conjunctions) and Stylistic (Osman-readability and type-Token-Ratio).

Using the complete set of features, J48 algorithm achieved an accuracy of 75.66%. After reducing the set of features we found that the top 8 features can achieve an accuracy of 75.02% using J48. This means that 16% of the features are enough to achieve a similar accuracy to that of using

[5] http://nlp.stanford.edu/projects/arabic.shtml

[6] Dialectal Arabic Tweets (DARTS) http://qufaculty.qu.edu.qa/telsayed/datasets/

[7] http://www.uncorpora.org/

[8] https://www.cs.waikato.ac.nz/ml/weka/

all the features combined. This 84% reduction helped in simplifying the model and shortening the training time.

5.2. Feature-Group Filtering

In order to examine the effect of each feature group (section 4.2.), we ran the classifiers using all of the training data testing on each feature-group individually and combined. This process helped us determine which set of features (feature-group) contribute most to identifying dialects. It also helped demonstrate which feature-groups are not meant to be used together, as such combination may increase complexity.

6. Results and Discussion

Overall, the best machine learning algorithm correctly distinguished dialects and MSA with more than 76% accuracy. This was calculated by training a SVM classifier using all set of features. We later show the detailed classification results after reducing the number of features as described in Section 5.. We compare that to our two baselines below.

6.1. Baselines

We used the most frequent class as our first baseline model. We also used a unigram, bigram and trigram word Naïve Bayes n-gram classifier as our second more intelligent baseline. Table 2 shows the accuracy scores of each baseline. The most frequent class (EGY) achieved an accuracy of 24% for the first baseline and reached 52% when using the second, more sensible baseline. The second score is still quite low due to the high bivalency between the dialects, especially in relation to short sentences, which in some cases were only one word long. On average, each instance contains 40 words with more than 3,000 instances containing less than 20 words. It is very difficult even for humans to guess the dialect for instances with one bivalent word such as "نعم" [yes], "رياضة" [sport] and "تعليم" [education], a task that is deemed impossible for a machine since these words are bivalent. Therefore, we expect our more refined, linguistically informed features will perform better to help the machine distinguish between the dialects.

Baseline	Accuracy
Most frequent class	24.62%
Word n-gram	52.07%

Table 2: Baseline Results

6.2. Training Results

We used a 10-fold cross validation to evaluate the models and avoid over-fitting. As shown in Section 4. we used four classifiers: J48, SVM, Naïve Bayes and KNN.

Table 3 shows the accuracy of the models on our training dataset. SVM achieved the highest accuracy with higher recall and precision compared to the other algorithms.

With more than 76% accuracy, our model performed better than a previous approach (Ali et al., 2015) which used a set of lexical and acoustic features to train a classifier using a dialectal dataset generated using an Arabic Automatic Speech Recognition (ASR) system.

M	A%	R	P	F
SVM	**76.29**	**0.76**	**0.79**	**0.78**
J48	75.66	0.76	0.75	0.76
KNN	62.72	0.63	0.62	0.63
NB	56.96	0.57	0.75	0.65

Table 3: Training results using all features
M: Model, A%: Accuracy, R: Recall, P: Precision, F: F-Measure

Table 4 shows the results of using our group-feature in addition to the combination of the 8 features (AttSel) selected using a Classifier Subset Evaluator as explained in Section 5.. As shown in the table, our SBP features play a vital role in helping the classifier identify dialects.

Feature	J48	SVM	NB	KNN
SBP + Gram	75.95	75.32	56.61	64.09
AttSel	75.02	69.96	65.45	69.26
SBP	74.97	71.01	59.19	72.49
Sty + SBP	74.55	69.55	59.66	69.03
Sty + Gram	51.50	54.21	41.47	47.02
Gram	50.56	52.56	40.47	46.39
Sty	45.59	31.22	33.82	42.41

Table 4: Examining Feature Groups (training)
Sty: Stylistics, SBP: Subtractive Bivalency Profiling, Gram: Grammatical, AttSel: Attribute Selection

The results show that our SBP method alone outperformed all the other features and that combining SBP with other features such as Grammatical and Stylistic features provides a small boost to accuracy. The selected features helped the classifier to distinguish between the dialects with a minimised error rate.

However, the results show that the Grammatical and Stylistic features alone did not perform better than the intelligent baseline. Comparing tables 3 and 4 shows that Naïve Bayes and KNN as well slightly improve the results when combining SBP with Stylistic features.

Table 5 shows the confusion matrix for the SVM classifier which achieved the highest scores as shown in Table 3. The confusion matrix shows that some dialects are harder to classify than others due to overlap with other dialects which reinforces our earlier informal observation that dialects use words interchangeably. The table shows the classifier to miss-classify EGY near equally between the other dialects when considering the number of pairwise mis-classified items. The table shows less confusion between GLF, LAV and NOR dialects. We believe this is due to the use of the SBP method which made it easier for the classifier to distinguish between dialects. It is not quite clear which dialect is closest to MSA but from the table we can observe that LAV is most likely to be miss-classified as MSA more than the other dialects in this corpus. Finally, it is important to note that our classifier has managed to equally distinguish between dialects whereas we can see in the confusion matrix that none of the dialects has been individually highly mis-classified except for MSA and EGY. This shows that the selected features are of good quality but that there is still scope for further improvement.

	EGY	GLF	LAV	MSA	NOR	Total
EGY	**3,371**	62	37	518	73	4,061
GLF	403	**1,573**	17	535	18	2,546
LAV	509	67	**1,231**	601	55	2,463
MSA	393	86	67	**3,047**	138	3,731
NOR	101	17	7	206	**3,362**	3,693
					Total	16,494

Table 5: SVM Classifier Confusion Matrix (based on Table 3)

We believe that our model can perform better with more refinement to the training data which could help decrease the impurity of the instances. To show how further refinement could help increase the accuracy of a classifier we trained a model using the same set of reduced features but this time only considering instances with more than or equal to 20 words. We reached this threshold by training a machine to increase the threshold with an increment of 1 for each classification iteration and to stop when the accuracy stops increasing or stalls. This has helped increase the accuracy of the SVM classifier to around 78%. This clearly shows how short sentences affect the classifier's accuracy. We will not use this classifier for testing as we may end up penalising the classifier when tested over short sentences. By refining the training data to only select instances with more than 20 words we intend to show how words overlap between dialects. Having these words out of context makes it difficult for a prediction model to infer the correct dialect even with more instances than what we have in our training data.

6.3. Unseen Testing Results

To further test the classifiers we used a separate unseen dataset. The source of the unseen testing data is similar to those of the training instances shown in Section 3. The randomly selected unseen data has never been used in training the classifiers and we use it to demonstrate how the classifier performs when tested on new data. Table 6 shows the distribution of the testing data.

Label	Sentences	Words
EGY	1,741	40,768
GLF	1,092	17,070
LAV	1,056	18,215
MSA	1,600	29,759
NOR	1,584	33,066
Total	7,073	138,878

Table 6: Testing data count

Table 7 shows the testing results using each classifier and set of features. In line with the training results (Section 6.) the testing results show that the SBP features alone outperformed all the other features. Moreover, combining SBP with other features such as Grammatical and Stylistic features slightly boosts classification accuracy. The testing results outperformed the two baselines in Section 4.1. which is also shown in the table when using n-gram features (c42%). The results show using n-gram features did not perform well on new unseen dataset. This could be due to the presence of new vocabulary items that the classifier has not encountered before. This suggests that our SBP feature group in addition to Grammatical and Stylistic features can still identify dialects fairly well even in the presence of vocabulary that the classifier has not seen before.

Feature	J48	SVM	NB	KNN
All	66.12	64.43	47.39	53.36
SBP + Gram	63.66	63.58	46.58	52.38
AttSel	60.61	63.96	56.47	56.78
SBP	57.06	61.73	47.65	57.83
Sty + SBP	57.04	60.17	49.39	54.47
Sty + Gram	55.04	50.52	40.63	44.19
Gram	52.49	50.38	38.92	42.17
Sty	44.60	45.08	37.30	40.19
n-gram	42.78	31.02	32.36	38.86

Table 7: Testing results
Sty: Stylistics, SBP: Subtractive Bivalency Profiling, Gram: Grammatical, AttSel: Attribute Selection, All: All features as in Table 3.

7. Conclusion

In this paper, we used machine learning to automatically detect dialects in a dataset comprising four Arabic dialects groups (Egyptian, Gulf, Levant and North African) in addition to Standard Arabic by applying a new method termed Subtractive Bivalency Profiling (SBP). The results showed that our SBP method alone outperformed all the other individual features and that the results improve slightly when combining SBP with other features. Code and other resources used in this paper are released freely on our GitHub repository.[9]

8. Acknowledgment

We would like to thank the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University for funding this work.

9. Bibliographical References

Abainia, K., Ouamour, S., and Sayoud, H. (2016). Effective language identification of forum texts based on statistical approaches. *Information Processing & Management*, 52(4):491 – 512.

Ali, A. M., Bell, P., and Renals, S. (2015). Automatic dialect detection in arabic broadcast speech. *Computing Research Repository*, abs/1509.06928.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Semitic '09, pages 53–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Blodgett, L. S., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case

[9]https://github.com/drelhaj/ArabicDialects

study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130. Association for Computational Linguistics.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Dunning, T. (1994). Statistical identification of language. Technical report, New Mexico State University.

El-Haj, M. and Rayson, P. (2016). Osman – a novel arabic readability metric. In *10th Language Resources and Evaluation Conference (LREC 2016)*, pages 250–255, Portoroz, Slovenia.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). Aida: Identifying code switching in informal arabic text. In *First Workshop on Computational Approaches to Code Switching. EMNLP 2014*, pages 94–101, Doha, Qatar.

Garcia-Barrero, D., Feria, M., and Turell, M. T. (2013). Using function words and punctuation marks in arabic forensic authorship attribution. In *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists*, pages 42–56, Porto, Portugal.

Gupta, Kumar, D., Kumar, S., and Ekbal, A. (2015). Machine learning approach for language identification & transliteration. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 60–64, New York, NY, USA. ACM.

Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. (2008). Guidelines for annotation of arabic dialectness. In *LREC Workshop on HLT and NLP within the Arabic World*, pages 49–53, Marrakesh, Morocco.

Hofland, K. and Johansson, S. (1982). *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Laboreiro, G., Bošnjak, M., Sarmento, L., Rodrigues, E. M., and Oliveira, E. (2013). Determining language variant in microblog messages. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 902–907, New York, NY, USA. ACM.

Lui, M. and Baldwin, T. (2014). Proceedings of the 5th workshop on language analysis for social media (lasm). pages 17–25. Association for Computational Linguistics.

Mejdell, G. (2011). Strategic bivalency in written 'mixed style'? a reading of ibrahim isa in al-dustur. In *Proceedings of the 9th International Arabic Dialectology Association (Aida) Conference*, Aida 9th, pages 273–278.

Ryding, K. (2014). *Arabic: A Linguistic Introduction*. Cambridge University Press.

Souter, C., Churcher, G., Hayes, J., Hughes, J., and Johnson, S. (1994). N-gram-based text categorization. In *Natural Language Identification using Corpus-based Models*, pages 161–175. Hermes Journal of Linguistics.

Woolard, K. A. and Genovese, E. N. (2007). Strategic bivalency in latin and spanish in early modern spain. *Language in Society*, 36(4):487–509.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Comput. Linguist.*, 40(1):171–202, March.