Comparing Approaches to the Identification of Similar Languages

Marcos Zampieri[1,2], Binyam Gebrekidan Gebre[3], Hernani Costa[4] and Josef van Genabith[1,2]
Saarland University, Germany[1]
German Research Center for Artificial Intelligence (DFKI), Germany[2]
Max Planck Computing and Data Facility, Germany[3]
University of Malaga, Spain[4]

Abstract

This paper describes the submission made by the MMS team to the Discriminating between Similar Languages (DSL) shared task 2015. We participated in the closed submission track using only the dataset provided by the shared task organisers which contained short texts from 13 similar languages and language varieties. We submitted three runs using different systems and compare their performance. As a result, our best system achieved 95.24% accuracy for test set A (containing original texts) and 92.78% accuracy for test set B (containing texts without named entities).

1 Introduction

Automatic language identification is an important task in Natural Language Processing (NLP), which consists of applying computational methods to identify the language a document is written in. Language identification is often modelled as a classification task and it is often the first processing stage of many NLP applications and pipelines. Although language identification is largely considered to be a solved task, recent studies have shown that language identification systems often fail to achieve satisfactory performance across different datasets and domains (Lui and Baldwin, 2011), particularly with: datasets containing short pieces of texts such as *tweets* (Zubiaga et al., 2014); code-switching data (Solorio et al., 2014); or when discriminating between very similar languages (Zampieri et al., 2014).

Given these challenges, the Discriminating between Similar Languages (DSL) shared task provides an excellent opportunity for researchers interested in evaluating and comparing their systems' performance on discriminating between similar languages and language varieties using short text excerpts extracted from journalistic texts. For this purpose, the MMS[1] team developed three systems for the closed submission track of the DSL shared task 2015. The systems are explained in more detail in Section 4.

The remainder of the paper is structured as follows. First, Section 2 presents the most relevant approaches in the field. The DSL shared task 2015 is described in detail in Section 3. Then, our approach and the results obtained are presented in Sections 4 and 5 Finally, Section 6 presents the final remarks and highlights our future plans for improving the systems.

2 Related Work

There have been a number of papers published about the identification or discrimination of similar languages in recent years. Most of them use supervised classification algorithms and words and characters as features to solve the task. Unlike general-purpose language identification, most of the systems trained to discriminate between similar languages perform best using high order character n-grams and word n-gram representations.

Different groups or pairs of similar languages and language varieties have been studied using data from different sources such as standard contemporary newspapers and social media. Recent studies include: Indian languages (Murthy and Kumar, 2006), Malay and Indonesian (Ranaivo-Malançon, 2006), Mainland, Singapore and Taiwanese Chinese (Huang and Lee, 2008), Brazilian and European Portuguese (Zampieri and Gebre, 2012), South Slavic languages (Tiedemann

[1]MMS is an acronym for our affiliations/locations (Malaga, Munich and Saarland). In the shared task report (Zampieri et al., 2015) the team is displayed as MMS*. The * indicates that a shared task organiser is a team member.

and Ljubešić, 2012; Ljubešić and Kranjčić, 2015) English varieties (Lui and Cook, 2013), Spanish varieties (Zampieri et al., 2013; Maier and Gómez-Rodrıguez, 2014), and Persian and Dari (Malmasi and Dras, 2015).

Over the last few years there has been a significant increase of interest in the computational processing of Arabic. This is evidenced by a number of research papers on different NLP tasks and applications including the identification/discrimination of Arabic dialects (Elfardy and Diab, 2014; Zaidan and Callison-Burch, 2014; Tillmann et al., 2014; Sadat et al., 2014; Salloum et al., 2014; Malmasi et al., 2015). From a purely engineering perspective, discriminating between dialects poses the same challenges as the discrimination between similar languages and language varieties.

3 The DSL Task

The shared task organisers provided all participants with an updated version of the DSL corpus collection v.2.0 (DSLCC) (Tan et al., 2014). This corpus is composed of 14 classes, 13 languages[2] and one class containing documents written in previously 'unseen' languages to emulate a real-world language identification scenario. Table 1 presents the languages included in the DSLCC v.2.0 corpus grouped by similarity.

Language/ Variety	Code
Bosnian	*bs*
Croatian	*hr*
Serbian	*sr*
Indonesian	*id*
Malay	*my*
Czech	*cz*
Slovak	*sk*
Brazilian Portuguese	*pt-BR*
European Portuguese	*pt-PT*
Argentine Spanish	*es-AR*
Castilian Spanish	*es-ES*
Macedonian	*bg*
Bulgarian	*mk*
Unknown	*xx*

Table 1: DSL corpus by language and variety.

In detail, the corpus collection contains 308,000 short text excerpts sampled from journalistic texts (22,000 per class) varying between 20 and 100 tokens per excerpt.

It is important to mention that these 22,000 texts per class are divided into 3 partitions, i.e. 18,000, 2,000 and 2,000 instances for training, development and testing, respectively. The test set is further subdivided into two test sets (A and B), each one containing 1,000 instances. While the test set A contains original texts, the organisers replaced named entities for place holders in the set B in order to decrease thematic bias in the classification process. Below we present an example of a Portuguese instance containing place holders *#NE#* instead of the named entities.

(1) Compara #NE# este sistema às indulgências vendidas pelo #NE# na #NE# #NE# quando os fiéis compravam a redenção das suas almas dando dinheiro aos padres.

Regarding the choice of only participating in the closed submission track, we first analysed the results of the 2014 edition where we realised that only two teams decided to participate in both open and closed submission tracks, namely UMich (King et al., 2014) and UniMelb-NLP (Lui et al., 2014). Both of them had better performance in the closed submission track and reported that more training data does not necessarily lead to higher performance and that the features learned by the classifiers are, to a certain extent, dataset specific. Therefore, we decided to use only the dataset provided by the organisers and only participate in the closed submission track.

4 Approach

Given that each team was allowed to submit a maximum of three runs to each track (closed and open), we decided to take this opportunity to test and compare different approaches. To do that, we developed three systems based on team MMS-member's previous work in language identification and related tasks. The first two systems were previously used for the Native Language Identification (NLI) (Gebre et al., 2013) and the third one has been applied to language variety identification. The following is a list of the three systems and the their corresponding submission runs:

- **Run 1** - Logistic Regression with TF-IDF Weighting

[2]For the sake of simplicity, we refer to both languages and language varieties as languages.

- **Run 2** - SVM with TF-IDF Weighting
- **Run 3** - Likelihood Estimation

It is important to mention that in each run we used different groups of features, all of them based on n-grams. In detail, for *Run 1* and *Run 2* we used n-grams ranging from bi- to seven-grams and 5-grams for *Run 3*.

4.1 TF-IDF Weighting

Term Frequency - Inverse Document Frequency (TF-IDF)[3] weighting measure was used in the systems developed for *Run 1* and *Run 2*.

Term Frequency refers to the number of times a particular term appears in a text.[4] It seems intuitive to think that a term that occurs more frequently tends to be a better identifier for the text than a term that occurs less frequently, however, this intuition does not take into account the relationship between the frequency of a term and its importance to the text. For this reason, we computed a logarithmic relationship (sublinear TF scaling) (Manning et al., 2008):

$$wf_{t,d} = \begin{cases} 1 + \log(tf_{t,d}) & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $wf_{t,e}$ refers to weight and $tf_{t,e}$ refers to the frequency of term t in document d.

The $wf_{t,d}$ weight represents the importance of a term in a document based on its frequency. However, not all terms that occur frequently in a text are equally important for our purpose. As an example, lets suppose we need to train a classifier to distinguish between British and American English varieties. Words like *the, of, and* will be very frequent, but they are not discriminative, mostly because they are frequent in both varieties. On the other hand, words like *London* or *rubbish* might not be as frequent as *the, of, and*, yet, they are better discriminative words for British English. Therefore, the actual importance of a term for this task depends on how infrequent the term is in other texts. This can be modelled using Inverse Document Frequency (IDF). IDF is based on the assumption that a term which occurs in many texts is not a good discriminator, and should be given less weight than one which occurs in fewer texts. To summarize, IDF is the *log* of the inverse probability of a term being found in any document (Salton and McGill, 1986):

$$idf(t_i) = \log \frac{N}{n_i} \quad (2)$$

where N is the number of documents in the corpus, and term t_i occurs in n_i of them.

TF gives more weight to a frequent term in a document whereas IDF decreases this weight if the term occurs in many documents. On their own, these measures are not very powerful as when combined together to form the well-known TF-IDF measure. The TF-IDF formula combines the weights of TF and IDF by multiplying them. Returning to our example, *the* is a frequent English word so its TF value will be high, however, it is a frequent word in all English texts, in turn making its IDF value low.

Equation 3 shows the final weight that each term in a document gets before normalisation.

$$w_{i,d} = (1 + \log(tf_{t,d})) \times \log \frac{N}{n_i} \quad (3)$$

The texts included in the shared task dataset have different lengths ranging between 20 and 100 tokens each. To cope with this variation we normalised each document feature vector to unit length so that document length does not severely impact term weights. The resulting document feature vectors are fed into two different classifiers, Logistic Regression and SVM.

4.2 Classifiers

Systems developed for *Run 1* and *Run 2* were previously used in the Native Language Identification (NLI) (Gebre et al., 2013) shared task 2013 (Tetreault et al., 2013) by the Cologne-Nijmegen team with good results. They both rely on the TF-IDF weighting scheme combined with two different classifiers.

For *Run 1*, we opt for Logistic Regression using the LIBLINEAR open source library (Fan et al., 2008) from scikit-learn (Pedregosa et al., 2011) and fix the regularisation parameter to 100.0. This regression algorithm has been used in different classification problems including for example temporal text classification (Niculae et al., 2014).

[3] The TF-IDF description presented in this section is based on our previous work (Gebre et al., 2013)

[4] In our experiments, terms are n-grams of characters, words, part-of-speech tags or any combination of them.

For *Run 2*, we used a Support Vector Machine classifier (Joachims, 1998). This approach delivered a slightly better performance than Logistic Regression during the NLI shared task. On a very challenging dataset containing TOEFL essays written by speakers of 11 different languages, TF-IDF with SVM reached 81.4% and 84.6% accuracy on the test set when using 10-fold cross validation.

Finally, for *Run 3* we use a simple, yet efficient and fast method that combines Laplace smoothing and a probabilistic classifier. The approach was previously applied to distinguish Brazilian and European Portuguese texts (Zampieri and Gebre, 2012) and it is available as an open source tool called *VarClass* (Zampieri and Gebre, 2014). The likelihood function is calculated as described in equation 1.

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \quad (4)$$

where N is the number of n-grams in the test text, n_i is the ith n-gram and L stands for the language models. Given a test text, we calculate the probability for each of the language models. The language model with the highest probability determines the identified language of the text.

5 Results

We start by reporting the official shared task results in terms of accuracy. Table 2 highlights the best results for each dataset.

Run	**Test Set A**	**Test Set B**
Run 1	94.09%	**92.77%**
Run 2	**95.24%**	**92.77%**
Run 3	94.07%	92.47%
Rank	2nd *out of 9*	4th *out of 7*

Table 2: Overall accuracy.

Results obtained by the three systems are very similar. Nevertheless, the SVM with TF-IDF Weighting approach obtained slightly better overall performance (*Run 2*). As we expected, the systems' performance drops from test set A to test set B. This means that our systems rely on named entities to discriminate between similar languages. It is important to point out that we did not do any specific training with the blinded named entities. Probably we could have achieved better results if we had prepared our systems to cope with this variation.

Table 3 presents the accuracy obtained by our best system (SVM with TF-IDF Weighting - *Run 2*) for each of the 14 classes. The results show that our best system achieved perfect performance in two of the language groups (Czech/ Slovak and Bulgarian/ Macedonian), probably due to exclusive characters present in one of the languages, as well as in identifying the 'unseen' languages in test set A.

Language/Variety	**Test Set A**	**Test Set B**
Bosnian	83.5%	76.6%
Croatian	91.8%	92.2%
Serbian	93.9%	90.7%
Indonesian	99.2%	97.5%
Malay	99.4%	99.5%
Czech	100%	99.9%
Slovak	100%	100%
Brazilian Portuguese	93.6%	90.5%
European Portuguese	93.0%	86.7%
Argentine Spanish	91.2%	89.2%
Castilian Spanish	94.8%	94.5%
Macedonian	100%	100%
Bulgarian	100%	100%
Unknown	100%	99.8%

Table 3: *Run 2*: performance per language.

Although the performance did not drop for Croatian and Malay when comparing test set A and B as it did for the rest of the languages, we do not think that this reflects any property of Croatian nor Malay nor any characteristics of the dataset. This is a simple preference of the classifier when distinguishing Croatian from Bosnian and Serbian, and Malay from Indonesian.

Tables 4, 5 and 6 present the confusion matrices obtained by the three systems using the 2,000 gold test instances.

Table 6 shows that Likelihood Estimation used for *Run 3* achieved higher scores when discriminating between language varieties, by classifying 1,912 Peninsular Spanish texts and 1,867 Brazilian Portuguese texts correctly. On the other hand, it was the only method which did not score 100% when classifying 'unseen' languages. Due to its simplicity, this method is well suited to discriminate between language varieties, hence the good results obtained in binary classification for Portuguese (Zampieri and Gebre, 2012), but

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1578	0	0	0	241	0	0	0	0	0	0	181	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1774	226	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	227	1773	0	0	0	0	0	0	0	0	0
hr	0	132	0	0	0	1841	0	0	0	0	0	0	26	1
id	0	0	0	0	0	0	1979	0	21	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	30	0	1970	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1826	174	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	222	1778	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1873	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

Table 4: Confusion Matrix *Run 1* - Axis Y represents the actual classes and Axis X the predicted classes.

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1661	0	0	0	193	0	0	0	0	0	0	146	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1796	204	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	209	1791	0	0	0	0	0	0	0	0	0
hr	0	135	0	0	0	1843	0	0	0	0	0	0	21	1
id	0	0	0	0	0	0	1988	0	12	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	19	0	1981	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1844	156	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	166	1834	0	0	0
sk	0	0	1	0	0	0	0	0	0	0	0	1999	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1891	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

Table 5: Confusion Matrix *Run 2* - Axis Y represents the actual classes and Axis X the predicted classes.

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1623	0	0	0	198	0	0	0	0	0	0	179	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1623	377	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	88	1912	0	0	0	0	0	0	0	0	0
hr	0	205	0	0	0	1746	0	0	0	0	0	0	49	0
id	0	0	0	0	0	0	1980	0	20	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	8	0	1992	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1867	133	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	236	1764	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	107	0	0	0	36	0	0	0	0	0	0	1857	0
xx	5	2	0	5	7	3	0	0	0	0	0	0	2	1976

Table 6: Confusion Matrix *Run 3* - Axis Y represents the actual classes and Axis X the predicted classes.

it clearly does not cope well with unseen data. Consequently, this method can be considered a good choice for situations in which all classes are known *a priori*.

6 Conclusion

This paper presented the MMS entry to the Discriminating between Similar Languages (DSL) shared task. We submitted three different approaches to deal with the task in hand, and their overall scores turned out to be very similar. The linear SVM classifier combined with TF-IDF weighting (*Run 2*) achieved slightly better results than the other two methods, i.e. 95.24% against 94.07% and 94.09% accuracy on test set A. The system ranked 2nd (out of 9 teams) on the test set A and 4th (out of 7 teams) on the test set B.

Based on the results, we observed that the systems' performance drop from test set A to test set B. This was already expected because named entities play an important role in this kind of task. One of the ways to cope with the influence of named entities in text classification is to use delexicalised text representations relying on POS tags or hybrid representations mixing word forms and grammatical categories. In our previous work, however, the results obtained using POS tags to discriminate between Spanish varieties, indicate that the use of more abstract text representations do not result in performance gain (Zampieri et al., 2013). In future work we would like to return to the question of text representation and investigate whether we can propose features that deliver high performance across multiple datasets.

An interesting approach would be to model these three systems hierarchically. This would result in a two-level classification task, first identifying the language group (grouped by similarity) and then the language itself. This approach was proposed by the NRC team, the DSL winner of the 2014 edition (Goutte et al., 2014). In the future we plan to investigate whether performing classification on two levels would increase the overall score or not.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471.

References

Heba Elfardy and Mona T Diab. 2014. Sentence level dialect identification in Arabic. In *Proceedings of ACL*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskens. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the 8th BEA workshop*, Atlanta, USA.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the VarDial Workshop*, Dublin, Ireland.

Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of PACLIC*, pages 404–410.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142. Springer.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the VarDial Workshop*, Dublin, Ireland.

Nikola Ljubešić and Denis Kranjčić. 2015. Discriminating between closely related languages on twitter. *Informatica*, 39(1).

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of Australasian Language Technology Workshop*, pages 5–15.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 2014. Exploring methods and resources for discriminating similar languages. In *Proceedings of VarDial*.

Wolfgang Maier and Carlos Gómez-Rodrıguez. 2014. Language variety identification in Spanish tweets. In *Proceedings of the LT4CloseLang Workshop*.

Shervin Malmasi and Mark Dras. 2015. Automatic Language Identification for Persian and Dari texts. In *Proceedings of PACLING 2015*, pages 59–64.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of PACLING 2015*, pages 209–217, Bali, Indonesia, May.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Kavi Narayana Murthy and G Bharadwaja Kumar. 2006. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(01):57–80.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of EACL*. ACL.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Bali Ranaivo-Malançon. 2006. Automatic identification of close languages - case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126–134.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of SocialNLP 2014*.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of ACL*, pages 772–778, Baltimore, USA.

Gerard Salton and Michael J McGill. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. 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.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of The Workshop on Building and Using Comparable Corpora (BUCC)*, Reykjavik, Iceland.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of BEA*, Atlanta, GA, USA, June.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. 2014. Improved sentence-level Arabic dialect classification. In *Proceedings of the VarDial Workshop*, pages 110–119, Dublin, Ireland, August.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS*, pages 233–237.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2014. Varclass: An open source language identification tool for language varieties. In *Proceedings of Language Resources and Evaluation (LREC)*.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN*, pages 580–587.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the VarDial Workshop*, pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the dsl shared task 2015. In *Proceedings of LT4VarDial*, Hissar, Bulgaria.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. 2014. Overview of tweetlid: Tweet language identification at sepln 2014. In *Proceedings of the Tweet Language Identification Workshop (TweetLID)*, Girona, Spain.