A Report on the DSL Shared Task 2014

Marcos Zampieri[1], Liling Tan[2], Nikola Ljubešić[3], Jörg Tiedemann[4]
Saarland University, Germany[1,2]
University of Zagreb, Croatia[3]
Uppsala University, Sweden[4]
marcos.zampieri@uni-saarland.de, liling.tan@uni-saarland.de
jorg.tiedemann@lingfil.uu.se, nljubesi@ffzg.hr

Abstract

This paper summarizes the methods, results and findings of the Discriminating between Similar Languages (DSL) shared task 2014. The shared task provided data from 13 different languages and varieties divided into 6 groups. Participants were required to train their systems to discriminate between languages on a training and development set containing 20,000 sentences from each language (closed submission) and/or any other dataset (open submission). One month later, a test set containing 1,000 unidentified instances per language was released for evaluation. The DSL shared task received 22 inscriptions and 8 final submissions. The best system obtained 95.7% average accuracy.

1 Introduction

Discriminating between similar languages is one of the bottlenecks of state-of-the-art language identification systems. Although in recent years systems have been trained to discriminate between more languages[1], they still struggle to discriminate between similar languages such as Croatian and Serbian or Malay and Indonesian.

From an NLP point of view, the difficulty systems face when discriminating between closely related languages is similar to the problem of discriminating between standard national language varieties (e.g. American English and British English or Brazilian Portuguese and European Portuguese), henceforth varieties. Recent studies show that language varieties can be discriminated automatically using words or characters as features (Zampieri and Gebre, 2012; Lui and Cook, 2013) . However, due to performance limitations, state-of-the-art general-purpose language identification systems do not distinguish texts from different national varieties, modelling pluricentric languages as unique classes.

To evaluate how state-of-the-art systems perform in identifying similar languages and varieties, we decided to organize the Discriminating between Similar Languages (DSL)[2] shared task. This shared task was organized within the scope of the workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) in the 2014 edition of COLING.

The motivation behind the DSL shared task is two-fold. Firstly, we have observed an increase of interest in the topic. This is reflected by a number of papers that have been published about this task in recent years starting with Ranaivo-Malançon (2006) for Malay and Indonesian and Ljubešić et al. (2007) for South Slavic languages. In the DSL shared task we tried to include (depending on the availability of data) languages that have been studied in previous experiments, such as Croatian, English, Indonesian, Malay, Portuguese and Spanish.

The second aspect that motivated us to organize this shared task is that, to our knowledge, no shared task focusing on the discrimination of similar languages has been organized previously. The most similar shared tasks to DSL are the DEFT 2010 shared task (Grouin et al., 2010), in which systems were required to classify French journalistic texts with respect to their geographical location as well as the

[1] Brown (2013) reports results on a system trained to recognize more than 1,100 languages

[2] http://corporavm.uni-koeln.de/vardial/sharedtask.html

decade in which they were published. Other related shared tasks include the ALTW 2010 multilingual language identification shared task, a general-purpose language identification task containing data from 74 languages (Baldwin and Lui, 2010) and finally the Native Language Identification (NLI) shared task (Tetreault et al., 2013) where participants were provided English essays written by foreign students of 11 different mother tongues (Blanchard et al., 2013). Participants had to train their systems to identify the native language of the writer of each text.

2 Related Work

Among the first studies to investigate the question of discriminating between similar languages is the study published by Ranaivo-Malançon (2006). The author presents a semi-supervised model to distinguish between Indonesian and Malay, two closely related languages from the Austronesian family represented in the DSL shared task. The study uses the frequency and rank of character trigrams derived from the most frequent words in each language, lists of exclusive words, and the format of numbers (Malay uses decimal points whereas Indonesian uses commas). The author compares the performance of this method with the performance obtained by *TextCat* (Cavnar and Trenkle, 1994).

Ljubešić et al. (2007) proposed a computational model for the identification of Croatian texts in comparison to Slovene and Serbian, reporting 99% recall and precision in three processing stages. The approach includes a 'black list', which increases the performance of the algorithm. Tiedemann and Ljubešić (2012) improved this method and applied it to Bosnian, Croatian and Serbian texts. The study reports significantly higher performance compared to general purpose language identification methods.

The methods applied to discriminate between texts from different language varieties and dialects are similar to those applied to similar languages[3]. One of the methods proposed to identify language varieties is by Huang and Lee (2008). This study presented a bag-of-words approach to classify Chinese texts from the mainland and Taiwan with results of up to 92% accuracy.

Another study that focused on language varieties is the one published by Zampieri and Gebre (2012). In this study, the authors proposed a log-likelihood estimation method along with Laplace smoothing to identify two varieties of Portuguese (Brazilian and European). Their approach was trained and tested in a binary setting using journalistic texts with accuracy results above 99.5% for character n-grams. The algorithm was later adapted to classify Spanish texts using not only the classical word and character n-grams but also POS distribution (Zampieri et al., 2013).

The aforementioned study by Lui and Cook (2013) investigates computational methods to discriminate between texts from three different English varieties (Canadian, Australian and British) across different domains. The authors state that the results obtained suggest that each variety contains characteristics that are consistent across multiple domains, which enables algorithms to distinguish them regardless of the data source.

Zaidan and Callison-Burch (2013) propose computational methods for the identification of Arabic language varieties[4] using character and word n-grams. The authors built their own dataset using crowdsourcing and investigated annotators' behaviour, agreement and performance when manually tagging instances with the correct label (variety).

3 Methods

In the following subsections we will describe the methodology adopted for the DSL shared task. Due to the lack of comparable resources, the first decision we had to take was to create a dataset that could be used in the shared task and also redistributed to be used in other experiments. We opted for the creation of a corpus collection based on existing datasets as discussed in 3.1 (Tan et al., 2014).

Groups interested in participating in the DSL shared task had to register themselves in the shared task website to receive the training and test data. Each group could participate in one or two types of

[3]In the DSL shared task and in this paper we did not distinguish between language varieties and similar languages. More on this discussion can be found in Clyne (1992) and Chamber and Trudgill (1998).

[4]Zaidan and Callison-Burch (2013) use the terms 'varieties' and 'dialects' interchangeably whereas Lui and Cook (2013) use the term 'national dialect' to refer to what previous work describes as 'national variety'.

submission as follows:

- **Closed Submission:** Using only the DSL corpus collection for training.
- **Open Submission:** Using any other dataset including or not the DSL collection for training.

In the open submission we did not make any distinction between systems using the DSL corpus collection and those that did not. This is different from the types of submissions for the NLI shared task 2013. The NLI shared task offered proposed two types of open submissions: open submission 1 - any dataset including the aforementioned TOEFL11 dataset (Blanchard et al., 2013) and open submission 2 - any dataset excluding TOEFL11.

For each of these submission types, participants were allowed to submit up to three runs, resulting in a maximum of six runs in total (three closed submissions and three open submissions).

3.1 Data

As previously mentioned, we decided to compile our own dataset for the shared task. The dataset was entitled DSL corpus collection and its compilation was motivated by the absence of a resource that allowed us to evaluate systems on discriminating similar languages. The methods behind the compilation of this collection and the preliminary baseline experiments are described in Tan et al. (2014).

The DSL corpus collection consists of 18,000 randomly sampled training sentences, 2,000 development sentences and 1,000 test sentences for each language (or variety) containing at least 20 tokens[5] each. The languages are presented in table 1 with their ISO 639-1 language codes[6]. For language varieties the country code is appended to the ISO code (e.g. *en-GB* refers to the British variety of English).

Group	Language/Variety	Code
A	Bosnian	*bs*
	Croatian	*hr*
	Serbian	*sr*
B	Indonesian	*id*
	Malay	*my*
C	Czech	*cz*
	Slovak	*sk*
D	Brazilian Portuguese	*pt-BR*
	European Portuguese	*pt-PT*
E	Argentine Spanish	*es-AR*
	Castilian Spanish	*es-ES*
F	British English	*en-GB*
	American English	*en-US*

Table 1: Language Groups - DSL 2014 Shared Task

For this collection, randomly sampled sentences from journalistic corpora (and corpora collections) were selected for each of the 13 classes. Journalistic corpora were preferred because they represent standard language, which is an important factor to be considered when working with language varieties. Other data sources (e.g. Wikipedia) do not make any distinction between language varieties and they are therefore not suitable for the purpose of the shared task. A number of studies mentioned in the related work section use journalistic texts for similar reasons (Huang and Lee, 2008; Grouin et al., 2010; Zampieri and Gebre, 2012)

Given what has been said in this section, we consider the collection to be a suitable comparable corpora from this task, which was compiled to avoid bias in classification towards source, register and topics. The

[5]We considered a token as orthographic units delimited by white spaces.

[6]`http://www.loc.gov/standards/iso639-2/php/English_list.php`

DSL corpus collection was distributed in tab delimited format; the first column contains a sentence in the language/variety, the second column states its group and the last column refers to its language code.[7]

3.1.1 Problems with Group F

There are no major problems to report regarding the organization of the shared task nor with the compilation of the DSL corpus collection apart from some issues in the Group F data. The organizers and a couple of teams participating in the shared task observed very poor performance when distinguishing instances from group F (British English - American British). For example, the baseline experiments described in Tan et al. (2014) report a very low 0.59 F-measure for Group F (the lowest score) and 0.84 for Group E (the second lowest score). Some of the teams asked human annotators to try to distinguish the sentences manually and they concluded that some instances were probably misclassified.

We decided to look more carefully at the data and noticed that the instances were originally tagged based on the websites (newspapers) that they were retrieved from and not the country of the original publication. There are, however, many cases of cross citation and republication of texts that the original data sources did not take into account (e.g. British texts that were later republished by an American website). As the DSL is a corpus collection and manually checking all 20,000 training and development instances per language was not feasible, we assumed that the original sources[8] from which the texts were retrieved provided the correct country of origin. The assumption was correct for all language groups but English.

To illustrate the issues above we present next some misclassified examples. Two particular cases raised by the UMich team are the following:

(1) I think they can afford to give North another innings and some time in Shield cricket and take another middle order batsman. (en-US)

(2) ATHENS, Ohio (AP) Albuquerque will continue its four-game series in Nashville Thursday night when it takes on the Sounds behind starter Les Walrond (3-4, 4.50) against Gary Glover, who is making his first Triple-A start after coming down from Milwaukee. (en-GB)

Example number one was tagged as American English because it was retrieved from the online edition of The New York Times but it was in fact first published in Australia. The second example is a text published by Associated Press describing an event that took place in Ohio, United States, but it was tagged as British English because it was retrieved by the UK Yahoo! sports section.

Our solution was to exclude the language group F from the final scores and perform a manual check in all its 1,000 test instances[9], thus giving the chance to participants to train their algorithms on other data sources (open submission).

3.2 Schedule

The DSL shared task spanned from March 20th when the training set was released, to June 6th when participants could submit a paper (up to 10 pages) describing their system. We provided one month between the release of the training and the test set. The schedule of the DSL shared task 2014 can be seen below.

Event	**Date**
Training Set Release	March 20th, 2014
Test Set Release	April 21st, 2014
Submissions Due	April 23rd, 2014
Results Announced	April 30th, 2014
Paper Submission	June 6th, 2014

Table 2: DSL 2014 Shared Task Schedule

[7] To obtain the data please visit: `https://bitbucket.org/alvations/dslsharedtask2014`

[8] See Tan et al. (2014) for a complete description of the data sources of the DSL corpus collection.

[9] Our manual check suggests that about 25% of the instances in the English dataset was likely to have been misclassified.

4 Results

This section summarises the results obtained by all participants of the shared task who submitted final results.[10] The DSL shared task included 22 enrolled teams from different countries (e.g. Australia, Estonia, Holland, Germany, United Kingdom and United States). From the 22 enrolled teams, eight of them submitted their final results. Most of the groups opted to exclusively use the DSL corpus collection and therefore participated solely in the closed submission track. Two of them compiled comparable datasets and also participated in the open submission.

Given that the dataset contained misclassified instances, group F (English) was not taken into account to compute the final shared task scores. In the next subsections we report results in terms of macro-average F-measure and accuracy.

4.1 Closed Submission

Table 3 presents the best F-measure and Accuracy results obtained by the eight teams that submitted their results for the closed submission track ordered by accuracy.

Team	Macro-avg F-score	Overall Accuracy
NRC-CNRC	0.957	0.957
RAE	0.947	0.947
UMich	0.932	0.932
UniMelb-NLP	0.918	0.918
QMUL	0.925	0.906
LIRA	0.721	0.766
UDE	0.657	0.681
CLCG	0.405	0.453

Table 3: Open Submission - Results

In the closed submissions, we observed a group of five teams whose systems (best runs) obtained results over 90% accuracy. This is comparable to what is described in the state-of-the-art literature for discriminating similar languages and language varieties (Tiedemann and Ljubešić, 2012; Lui and Cook, 2013). These five teams submitted system descriptions that allowed us to look in more detail at successful approaches for this task. System descriptions will be discussed in section 5.

Three of the eight teams obtained substantially lower scores, from 45.33% to 76.64% accuracy. These three groups unfortunately did not submit system description papers. From our point of view, this would create an interesting opportunity to look more carefully at the weaknesses of approaches that did not obtain good results in this task.

4.2 Open Submission

Only two systems submitted results for the open submission track and their F-measure and Accuracy results are presented in table 3.

Team	Macro-avg F-score	Overall Accuracy
UniMelb-NLP	0.878	0.880
UMich	0.858	0.859

Table 4: Closed Submission - Results

The UniMelb-NLP (Lui et al., 2014) group used data from different corpora such as the BNC, EUROPARL and Open Subtitles whereas UMich (King et al., 2014) compiled journalistic corpora from different sizes for each language ranging from 695,597 tokens for Malay to 20,288,294 tokens for British English.

[10] Visit `https://bitbucket.org/alvations/dslsharedtask2014/downloads/dsl-results.html` for more detail on the shared task results or at the aforementioned DSL shared task website.

Comparing the results of the closed to the open submissions, we observed that the UniMelb-NLP submission was outperformed by UMich system by about 1.5% accuracy in the closed submission, but in the open submission they scored 2.1% better than UMich. This difference can be explained by investigating these two factors: 1) the quality and amount of the collected training data; 2) the robustness of the method to obtain correct predictions across different datasets and domains as previously discussed by Lui and Cook (2013) for English varieties.

4.3 Accuracy per Language Group

In this subsection we look more carefully at the performance of systems in discriminating each class within groups A to E. Table 5 presents the accuracy scores obtained per language group for each team sorted alphabetically. The best score per group is displayed in bold.

	CLCG	LIRA	NRC-CNRC	QMUL	RAE	UDE	UMich	UniMelb-NLP
A	0.338	0.333	**0.936**	0.879	0.919	0.785	0.919	0.915
B	0.503	0.982	**0.996**	0.935	0.994	0.892	0.992	0.972
C	0.500	**1.000**	**1.000**	0.962	**1.000**	0.493	0.999	**1.000**
D	0.496	0.892	**0.956**	0.905	0.948	0.493	0.926	0.896
E	0.503	0.843	**0.910**	0.865	0.888	0.694	0.876	0.807

Table 5: Language Groups A to E - Accuracy Results

The top 5 systems plus the LIRA team obtained very good results for groups B (Malay and Indonesian) and C (Czech and Slovak). Four out of eight systems obtained perfect performance when discriminating Czech and Slovak texts. Perfect performance was not achieved by any of the systems when distinguishing Malay from Indonesian texts, but even so, results were fairly high and the best result was 99.6% accuracy obtained by the NRC-CNRC group. The perfect results obtained by four groups when distinguishing texts from group C suggest that Czech and Slovak texts are not as similar as we assumed before the shared task, and that they therefore possess strong systemic and/or orthographic differences that allow well-trained classifiers to perform perfectly. Figure 1 presents the accuracy results of the top 5 groups.

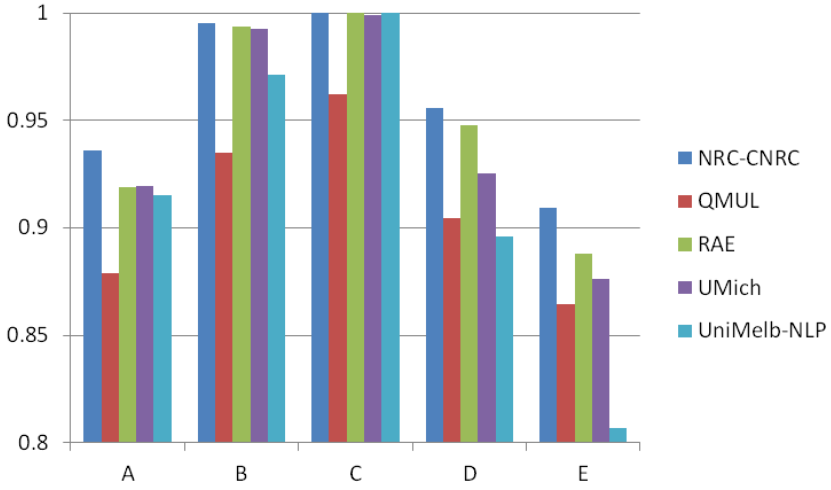

Figure 1: Language Groups A to E Accuracy - Top 5 Systems

Distinguishing between languages from group A (Bosnian, Croatian and Serbian), the only group containing 3 languages, proved to be a challenging task as discussed in previous research (Ljubešić et al., 2007; Tiedemann and Ljubešić, 2012). The best result was again obtained by the NRC-CNRC group with 93.5% accuracy. The groups containing texts written in different language varieties, namely D (Portuguese) and E (Spanish) were the most difficult to discriminate, particularly the Spanish varieties. These results also corroborate the findings of previous studies (Zampieri et al., 2013).

The QMUL system that was the 5th best system in the closed submission track did not outperform any of the other top 5 systems in groups A, B or C. However, the system did better when distinguishing texts from the two most difficult language groups (D and E), outperforming the UniMelb-NLP submission on two occasions. The simplicity of the approach proposed by the QMUL, which the author describes as 'a simple baseline' (Purver, 2014) may be an explanation for the regular performance across different language groups.

4.4 Results Group F

To document the problems in the group F (British and American English) dataset we included the results of both the open and closed submissions for this language group in table 6. As previously mentioned, submitting group F results was optional and we did not include these results in the final shared task results. Six out of eight systems decided to submit their predictions as closed submissions and the two groups participating in the open submission track also submitted their group F results.

Team	F-score	Accuracy	Type
UMich	0.639	0.639	Open
UniMelb- NLP	0.581	0.583	Open
NRC-CNRC	0.522	0.524	Closed
LIRA	0.450	0.493	Closed
RAE	0.451	0.481	Closed
UMich	0.463	0.464	Closed
UDE	0.451	0.451	Closed
UniMelb-NLP	0.435	0.435	Closed

Table 6: Group F - Accuracy Results

The results confirm the problems in the DSL dataset discussed in section 3.1.1. After a careful manual check of the 1,000 test instances, open submissions scores were still substantially lower than the other groups: 69.9% and 58.3% accuracy. Closed submissions proved to be impossible and only one of the six systems scored slightly above the 50% baseline.

It should be investigated more carefully in future research whether the poor results for group F reflect only the problems in the dataset or also the actual difficulty in discriminating between these two varieties of English. Moderate differences in orthography (e.g. *neighbour* (UK) and *neighbor* (US)) as well as lexical choices (e.g. *rubbish* (UK) and *garbage* (US) or *trousers* (UK) and *pants* (US)) are present in texts from these two varieties and these can be informative features for algorithms to discriminate between them. Discriminating between other English varieties already proved to be a challenging yet feasible task in previous research (Lui and Cook, 2013).

5 System Descriptions

All eight systems that submitted their final results to the shared task were invited to submit papers describing their systems and the top 5 systems in the closed track submitted their papers, namely: NRC-CNRC, RAE, UMich, UniMelb-NLP and QMUL.

The best scores were obtained by the NRC-CNRC (Goutte et al., 2014) team which proposed a two-step approach to predict first the language group than the language of each instance. The language group was predicted in a 6-way classification using a probabilistic model similar to a Naive Bayes classifier, and later the method applied SVM classifiers to discriminate within each group: binary for groups B-F and one versus all for group A, which contains three classes (Bosnian, Croatian and Serbian).

An interesting contribution proposed by the RAE team (Porta and Sancho, 2014) are the so-called 'white lists' inspired by the 'blacklist' classifier (Tiedemann and Ljubešić, 2012). These lists are word lists exclusive to a language or variety, similar to one of the features that Ranaivo-Malançon (2006) proposed to discriminate between Malay and Indonesian.

Two groups used Information Gain (IG) to select the best features for classification, namely UMich (King et al., 2014) and UniMelb-NLP (Lui et al., 2014). These teams were also the only ones to submit open submissions. The UniMelb-NLP team tried different classification methods and features (including delexicalized models) in each run. The best results were obtained by their own method, the off-the-shelf general-purpose language identification software *langid.py* (Lui and Baldwin, 2012). This method has been widely used for general-purpose language identification and its performance is regarded superior to similar general-purpose methods such as *TextCat*. In the shared task, the system was modelled hierarchically firstly identifying the language group that a sentence belongs to and subsequently the specific language, achieving performance comparable to the state-of-the-art, but still slightly below the other three systems.

The QMUL team (Purver, 2014) proposed a linear SVM classifier using words and characters as features. The author investigated the influence of the cost parameter c (from 1.0 to 100.0), in the classifiers' performance. The cost parameter c is responsible for the trade-off between maximum margin and classification errors. According to the system description the optimal parameter for this task lies between 30.0 and 50.0. Purver (2014) also notes that the linear SVM classifier performs well with word uni-gram language models in comparison to methods using character n-grams. This observation corroborates the findings of previous experiments that rely on words as important features to distinguish similar languages and varieties (Huang and Lee, 2008; Zampieri, 2013)

The features and algorithms presented so far, as well as the system paper descriptions, are summarised in table7.[11]

Team	**Algorithm**	**Features**	**System Paper**
NRC-CNRC	Prob. Class. and Linear SVM	Words 1-2, Char. 2-6	(Goutte et al., 2014)
RAE	MaxEnt	Words 1-2, Char. 1-5, 'Whitelist'	(Porta and Sancho, 2014)
UMich	Naive Bayes	Words 1-2, Char. 2-6 (IG Feat. Selection)	(King et al., 2014)
UniMelb-NLP	*langid.py*	Words, Char., POS (IG Feat. Selection)	(Lui et al., 2014)
QMUL	Linear SVM	Words 1, Char. 1-3	(Purver, 2014)

Table 7: Top 5 Systems - Features and Algorithms at a Glance

6 Conclusion

Shared tasks are an interesting way of comparing algorithms, computational methods and features using the same dataset. Given what has been presented in this paper, we believe that the DSL shared task filled an important gap in language identification and will allow other researchers to look in more detail at the problem of discriminating similar languages. Accurate methods for discriminating similar languages can help to improve performance not only in language identification but also in a number of NLP tasks and applications such as part-of-speech-tagging, spell checking and machine translation.

The best system obtained 95.71% accuracy and F-measure for a set of 11 languages and varieties divided into 5 groups (A to E), using only the DSL corpus collection. Systems that performed best modelled their algorithms to perform two-step predictions: first the language group, then the actual class and used characters and words as features. As we regard the corpus to be a balanced sample of the news domain, the results obtained confirm the assumption that similar languages and varieties possess systemic characteristics that can be modelled by algorithms in order to distinguish languages from other similar languages or varieties using lexical or orthographical features.

Another lesson learned from this shared task is regarding the compilation of group F (English) data. Researchers, including us, often rely on previously annotated meta-data which sometimes may contain inaccurate information and errors. Corpus collection for this purpose should be thoroughly checked (manually if possible). The issues with the group F might have discouraged some of the participants to continue in the shared task (particularly those who were interested only in the discrimination of English varieties).

[11] UniMelb-NLP experimented different methods in their 6 runs. In this report we commented on the algorithm that achieved the best performance.

6.1 Future Perspectives

The shared task was a very fruitful and positive experience for the organizers. We would like to organize a second edition of the shared task containing, for example, new language groups for which we could not find suitable corpora before the 2014 edition. This includes, most notably, the cases of Dutch and Flemish or the varieties of French and German which could not be included in the DSL shared task due to the lack of available data.

The DSL corpus collection is freely available and can be used as a gold standard for language identification or to train algorithms for other NLP tasks involving similar languages. We would like to use the dataset to investigate, for example, lexical variation between similar languages and varieties as proposed by Piersman et al. (2010) and Soares da Silva (2010) or syntactic variation using annotated data as discussed in Anstein (2013).

At present, we are investigating the influence of the length of texts in the discrimination of similar languages. It is a well known fact that the longer texts are, the more likely they are to contain features that allow algorithms to identify their language. However, this variable was not explored within the scope of the DSL shared task and we are using the DSL dataset and the results for this purpose. Another direction that our work may take is the linguistic analysis of the most informative features in classification as was done recently by Diwersy et al. (2014).

Acknowledgements

The authors would like to thank all participants of the DSL shared task for their comments and suggestions throughout the organization of this shared task. We would also like to thank Joel Tetreault and Binyam Gebrekidan Gebre for their valuable feedback on this report.

References

Stefanie Anstein. 2013. *Computational approaches to the comparison of regional variety corpora : prototyping a semi-automatic system for German*. Ph.D. thesis, University of Stuttgart.

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

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.

Ralf Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. In *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI 8082)*, pages 519–526, Pilsen, Czech Republic. Springer.

William Cavnar and John Trenkle. 1994. N-gram-based text catogorization. *3rd Symposium on Document Analysis and Information Retrieval (SDAIR-94)*.

Jack Chambers and Peter Trudgill. 1998. *Dialectology (2nd Edition)*. Cambridge University Press.

Michael Clyne. 1992. *Pluricentric Languages: Different Norms in Different Nations*. CRC Press.

Sascha Diwersy, Stefan Evert, and Stella Neumann. 2014. A semi-supervised multivariate approach to the study of language variation. *Linguistic Variation in Text and Speech, within and across Languages.*

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland.

Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et résultats du défi fouille de texte DEFT2010 où et quand un article de presse a-t-il été écrit? *Actes du sixième DÉfi Fouille de Textes*.

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 2008*, pages 404–410.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland.

Nikola Ljubešić, Nives Mikelic, and Damir Boras. 2007. Language identification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Meeting of the ACL*.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of Australasian Language Tchnology 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 the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland.

Yves Piersman, Dirk Geeraerts, and Dirk Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16:469–491.

Jordi Porta and José-Luis Sancho. 2014. Using maximum entropy models to discriminate between similar languages and varieties. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland.

Matthew Purver. 2014. A simple baseline for discriminating similar language. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland.

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

Augusto Soares da Silva. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese: endo/exogeneousness and foreign and normative influence. *Advances in Cognitive Sociolinguistics*.

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 the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, GA, USA, June. Association for Computational Linguistics.

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

Omar F Zaidan and Chris Callison-Burch. 2013. Arabic dialect identification. *Computational Linguistics*.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, pages 233–237, Vienna, Austria.

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 TALN2013*, pages 580–587, Sable d'Olonne, France.

Marcos Zampieri. 2013. Using bag-of-words to distinguish similar languages: How efficient are they? In *Proceedings of the 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI2013)*, pages 37–41, Budapest, Hungary.