Overview of the DSL Shared Task 2015

Marcos Zampieri[1,2], Liling Tan[1], Nikola Ljubešić[3], Jörg Tiedemann[4], Preslav Nakov[5]
Saarland University, Germany[1]
German Research Center for Artificial Intelligence (DFKI), Germany[2]
University of Zagreb, Croatia[3]
University of Helsinki, Finland[4]
Qatar Computing Research Institute, HBKU, Qatar[5]
marcos.zampieri@uni-saarland.de, liling.tan@uni-saarland.de
jorg.tiedemann@lingfil.uu.se, nljubesi@ffzg.hr, pnakov@qf.org.qa

Abstract

We present the results of the 2nd edition of the Discriminating between Similar Languages (DSL) shared task, which was organized as part of the LT4VarDial'2015 workshop and focused on the identification of very similar languages and language varieties. Unlike in the 2014 edition, in 2015 we had an *Others* category with languages that were not seen on training. Moreover, we had two test datasets: one using the original texts (test set A), and one with named entities replaced by placeholders (test set B). Ten teams participated in the task, and the best-performing system achieved 95.54% average accuracy on test set A, and 94.01% on test set B.

1 Introduction

Identifying the language of an input text is an important step for many natural language processing (NLP) applications, especially when processing speech or social media messages. State-of-the-art language identification systems perform very well when discriminating between unrelated languages on standard datasets. For example, Simões et al. (2014) used TED talks and reported 97% accuracy for discriminating between 25 languages. Yet, this is not a solved problem, and there are a number of scenarios in which language identification has proven to be a very challenging task, especially in the case of very closely-related languages. For example, despite their good overall results, Simões et al. (2014) had really hard time discriminating between Brazilian and European Portuguese, which has made them propose to "remove the Brazilian Portuguese and/or merge it with the European Portuguese variant" to increase system's performance.

So far, researchers in language identification have focused on the following challenges:

- **Increasing the coverage** of language identification systems by extending the number of languages that are recognizable, e.g., Xia et al. (2010) trained a system to identify over 1,000 languages, whereas Brown (2014) developed a language identification tool able to discriminate between over 1,300 languages.

- **Improving the robustness** of language identification systems, e.g., by training on multiple domains and various text types (Lui and Baldwin, 2011).

- **Handling non-standard texts**, e.g., very short (Zubiaga et al., 2014) or involving code-switching (Solorio et al., 2014).

- **Discriminating between very similar languages** (Tiedemann and Ljubešić, 2012), language varieties (Zampieri et al., 2014), and dialects (Sadat et al., 2014; Malmasi et al., 2015).

It has been argued that the latter challenge is one of the main bottlenecks for state-of-the-art language identification systems (Tiedemann and Ljubešić, 2012). Thus, this was the task that we focused on in our shared task on Discriminating between Similar Languages (DSL), which we organized as part of the LT4VarDial'2015 workshop at RANLP'2015.

This is the second edition of the task. The attention received from the research community and the feedback provided by the participants of the first edition motivated us to organize this second DSL shared task, where we made two important changes compared to the first edition.

First, in order to simulate a real-world language identification scenario, we included in the testing dataset some languages that were not present in the training dataset. Moreover, we included a second test set, where we substituted the named entities with placeholders to make the task more challenging and less dependent on the text topic and domain.

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 describes the general setup of the task, Section 4 presents the results of the competition, Section 5 summarizes the approaches used by the participants, and Section 6 offers conclusions.

2 Related Work

Language identification has attracted a lot of research attention in recent years, covering a number of similar languages and language varieties such as Malay and Indonesian (Ranaivo-Malançon, 2006), Persian and Dari (Malmasi and Dras, 2015a), Brazilian and European Portuguese (Zampieri and Gebre, 2012), varieties of Mandarin in China, Taiwan and Singapore (Huang and Lee, 2008), and English varieties (Lui and Cook, 2013), among others. This interest has eventually given rise to special shared tasks, which allowed researchers to compare and benchmark various approaches on common standard datasets. Below we will describe some of these shared tasks, including the first edition of the DSL task.

2.1 Related Shared Tasks

There have been a number of language identification shared tasks in recent years. Some were more general, such as the ALTW language identification shared task (Baldwin and Lui, 2010), while others focused on specific datasets or languages. Yet, the DSL shared task is unique as it is the only one to focus specifically on discriminating between *similar languages and language varieties*, providing a standardized dataset for this purpose.

The most closely-related shared task is the DEFT 2010 shared task (Grouin et al., 2010), which targeted language variety identification. However, it focused on French language varieties only, namely on texts from Canada and France. Moreover, it featured a temporal aspect, asking participants to identify *when* a given text was written. This aspect is not part of our DSL shared task, as we focus on contemporary texts.

Another popular research direction has been on language identification on Twitter, which was driven by interest in geolocation prediction for end-user applications (Ljubešić and Kranjčić, 2015). This interest has given rise to the Tweet-LID shared task (Zubiaga et al., 2014), which asked participants to recognize the language of tweet messages, focusing on English and on languages spoken on the Iberian peninsula such as Basque, Catalan, Spanish, and Portuguese. The Shared Task on Language Identification in Code-Switched Data held in 2014 (Solorio et al., 2014) is another related competition, where the focus was on tweets in which users were mixing two or more languages in the same tweet.

2.2 The First Edition of the DSL Task

For the first edition of the task, we compiled the *DSL Corpus Collection* (Tan et al., 2014), or DSLCC v.1.0, which included excerpts from journalistic texts from sources such as the SETimes Corpus[1] (Tyers and Alperen, 2010), HC Corpora[2] and the Leipzig Corpora Collection (Biemann et al., 2007), written in thirteen languages divided into the following six groups: Group A (Bosnian, Croatian, Serbian), Group B (Indonesian, Malay), Group C (Czech, Slovak), Group D (Brazilian Portuguese, European Portuguese), Group E (Peninsular Spanish, Argentine Spanish), and Group F (American English, British English).

In 2014, eight teams built systems and submitted results to the DSL language identification shared task (eight teams participated in the closed and two teams took part in the open condition), and five participants wrote system description papers. The results are summarized in Table 1, where the best-performing submissions, in terms of testing accuracy, are shown in bold.

Team	Closed	Open
NRC-CNRC	**0.957**	-
RAE	0.947	-
UMich	0.932	0.859
UniMelb-NLP	0.918	**0.880**
QMUL	0.906	-
LIRA	0.766	-
UDE	0.681	-
CLCG	0.453	-

Table 1: DSL 2014 results: accuracy.

[1] Published as part of OPUS (Tiedemann, 2012).

[2] `http://www.corpora.heliohost.org/`

The best accuracy in the closed submission track of the 2014 edition of the DSL shared task was achieved by the NRC-CNRC (Goutte et al., 2014) team, which used a two-step classification approach: they first made a prediction about the language group the target text might belong to, and then they selected a language from that language group. Members of this team participated again in 2015 under the name NRC.

The RAE team (Porta and Sancho, 2014) used 'white lists' of words that are used exclusively in a particular language or language variety.

The QMUL team (Purver, 2014) used a linear support vector machines (SVM) classifier with words and characters as features. They further paid special attention to the influence of the cost parameter c on the classifier's performance; this SVM parameter is responsible for the trade-off between maximum margin and classification errors at training time.

Two other participating teams, UMich (King et al., 2014) and UniMelb-NLP (Lui et al., 2014), used Information Gain as a selection criterion (Yang and Pedersen, 1997) to select a subset of features, trying to improve classification accuracy. The UniMelb-NLP team experimented with different classifiers and features, and eventually obtained their best results using their own software, *langid.py* (Lui and Baldwin, 2012).

The UMich and UniMelb-NLP teams compiled and used additional training resources and were the only teams to submit open submissions. However, the performance of these open submissions were worse than what they achieved in their closed submissions: accuracy dropped from 93.2% to 85.9% for UMich, and from 91.8% to 88.0% for UniMelb-NLP.

This worse performance of the open submissions was quite surprising. We had a closer look, and we hypothesized that this could be due to the abundance of named entities in our datasets. For example, participating systems could learn that a text that talks about Brazilian places, companies, politicians, etc. is likely to be in Brazilian Portuguese. These are legitimate features, but they are about the topic of the text and do not reflect linguistic characteristics, which we were hoping participants would focus on. Thus, in the 2015 edition of the task, we created two test sets, one containing the original texts, and one where we substituted the named entities with placeholders.

3 Task Setup

In this section, we describe the general setup of the DSL 2015 shared and unshared task tracks, the changes in v2.0 of the DSLCC dataset compared to v1.0, and the task schedule.

3.1 The Shared Task Track

The setup of the 2015 DSL *Shared Task* is similar to the one for the 2014 edition. However, we created a new updated v2.0 of DSLCC (Tan et al., 2014), extending it with new languages. We provided participants with standard splits into training and development subsets, and we further prepared two test sets, as described in Section 3.3 below. As in 2014, teams could make two types of submissions (for each team, we allowed up to three runs per submission type; in the official ranking, we included the run with the highest score only):

- **Closed submission:** Using only the DSLCC v2.0 for training.
- **Open submission:** Using any dataset other than DSLCC v2.0 for training.[3]

3.2 The Unshared Task Track

Along with the Shared Task, this year we proposed an *Unshared Task* track inspired by the unshared task in PoliInformatics held in 2014 (Smith et al., 2014). For this track, teams were allowed to use any version of DSLCC to investigate differences between similar languages and language varieties using NLP methods. We were interested in studying questions like these:

- Are there fundamental grammatical differences in a language group?
- What are the most distinctive lexical choices for each language?
- Which text representation is most suitable to investigate language variation?
- What is the impact of lexical and grammatical variation on NLP applications?

Although eleven teams subscribed for the Unshared Task track, none of them ended up submitting a paper for it. Therefore, below we will only discuss the Shared Task track.

Language/Variety	ISO 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*
Bulgarian	*bg*
Macedonian	*mk*
Others	*xx*

Table 2: DSLCC v2.0: the languages included in the corpus grouped by similarity. *Others* is a mixture of Catalan, Russian, Slovene, and Tagalog.

3.3 The DSLCC v2.0 Dataset

Version 2.0 of DSLCC (Tan et al., 2014) contains a total of 308,000 examples divided into fourteen language classes with 22,000 examples per class. Each example is a short text excerpt of 20–100 tokens,[4] sampled from journalistic texts colected from the same sources as in DSLCC v1.0. The fourteen classes are shown in Table 2; they represent thirteen languages and language varieties and one mixed class with documents written in four other languages, namely: Catalan, Russian, Slovene, and Tagalog.[5] We included the mixed Others class in order to emulate a real-world language identification scenario in which 'unknown' but similar languages might appear, thus making the task more challenging.

We partitioned the 22,000 examples for each language class into three parts as follows: 18,000 examples for training, 2,000 for development, and 2,000 for testing. We then further subdivided each test set into two test sets, A and B, each containing 1,000 instances per language. We kept the texts in test set A unchanged, but we preprocessed those in test set B by replacing all named entities with placeholders.[6]

We substituted the named entities with placeholders in order to avoid topic bias in classification and to evaluate the extent to which proper names can influence classifiers' performance.

As an example, here we show a Portuguese and a Spanish text: first the original texts, then versions thereof with named entities substituted by placeholders *#NE#*.

(1) Rui Nobre dos Santos explica que "a empresa pretende começar a exportar para Angola e Moçambique, em 2010", objectivo que está traçado desde 2007 "mas que ainda não foi possível concretizar", e aumentar as exportações para o Brasil.

(2) El jueves pasado se conoció que Schoklender había renunciado a su cargo, según la prensa local por una pelea con su hermano, que también trabaja en la entidad, al parecer por desacuerdos en el manejo de los fondos para la construcción de viviendas populares.

(3) 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.

(4) La cinta, que hoy se estrena en nuestro país, competirá contra #NE# la #NE#, de #NE#, #NE#, de #NE#, #NE#, de #NE# á, #NE# above all, de #NE#, y con la ganadora del #NE# de #NE#, #NE# A #NE# #NE#, de #NE#.

3.4 Shared Task Schedule

The second DSL shared task was open for two months, spanning from May 20, 2015, when the training data was released, to July 20, 2015, when the paper submissions were due. Teams had just over a month to train their systems before the release of the test data. The schedule of the DSL shared task 2015 is shown in Table 3.

Event	Date
Training set released	May 20, 2015
Test set released	June 22, 2015
Submissions due	June 24, 2015
Results announced	June 26, 2015
Paper submissions due	July 20, 2015

Table 3: The DSL 2015 Shared Task schedule.

[3] Training on DSLCC v1.0 also makes a submission open.

[4] In DSLCC v1.0, texts could be longer than 100 tokens.

[5] For the Unshared Task track, we further made available DSLCC v2.1, which extended DSLCC v2.0 with Mexican Spanish and Macanese Portuguese data.

[6] The script we used to substitute named entities with placeholders is available here: `https://github.com/Simdiva/DSL-Task/blob/master/blindNE.py`

Team	Closed (Normal)	Closed (No NEs)	Open (Normal)	Open (No NEs)	System Description Paper
BOICEV	✓	✓	-	-	(Bobicev, 2015)
BRUNIBP	✓	-	-	-	(Ács et al., 2015)
INRIA	✓	-	-	-	-
MAC	✓	✓	-	-	(Malmasi and Dras, 2015b)
MMS*	✓	✓	-	-	(Zampieri et al., 2015)
NLEL	✓	✓	✓	✓	(Fabra-Boluda et al., 2015)
NRC	✓	✓	✓	✓	(Goutte and Léger, 2015)
OSEVAL	-	-	✓	✓	-
PRHLT	✓	✓	-	-	(Franco-Salvador et al., 2015)
SUKI	✓	✓	-	-	(Jauhiainen et al., 2015a)
Total	**9**	**7**	**3**	**3**	**8**

Table 4: The participating teams in the DSL 2015 Shared Task.

4 Results

In this section, we present the results of the 2nd edition of the DSL shared task.[7] Most of the participating teams used DSLCC v2.0 only, and thus took part in the closed submission track. Yet, three of the teams collected additional data or used DSLCC v1.0, and thereby participated in the open submission.

4.1 Submitted Runs

A total of 24 teams subscribed to participate in the shared task, 10 of them submitted official runs, and 8 of the latter also wrote system description papers. These numbers represent a slight increase in participation compared to the 2014 edition, which attracted 22 teams, 8 submissions, and 5 system description papers.

Table 4 gives information about the ten teams that submitted runs, indicating the tracks they participated in. The table also includes references to their system description papers, when applicable. As one of the members of the MMS team was a shared task organizer, we have decided to mark the team with a star; and we do so in all tables. Still, this team did not have any unfair advantage, and competed under the same conditions as the rest.

4.2 Closed Submission

As in 2014, most teams chose to participate in the closed submission: 9 out of 10. All these 9 teams submitted runs for test set A, and their results are shown in Table 5. We can see that the best result was 95.54% accuracy, achieved by the MAC team, followed very closely by MMS and NRC, which both achieved 95.24% accuracy.

Rank	Team	Accuracy
1	MAC	95.54
2-3	MMS*	95.24
2-3	NRC	95.24
4	SUKI	94.67
5	BOBICEV	94.14
6	BRUNIBP	93.66
7	PRHLT	92.74
8	INRIA	83.91
9	NLEL	64.04

Table 5: Closed submission results for test set A.

Seven of the nine teams who took part in the open submission submitted runs for test set B; the results are shown in Table 6. We can see a drop in accuracy, which is to be expected. Once again, the MAC team performed best with 94.01% accuracy, followed by SUKI and NRC with 93.02% and 93.01%, respectively.

Rank	Team	Accuracy
1	MAC	94.01
2	SUKI	93.02
3	NRC	93.01
4	MMS*	92.78
5	BOBICEV	92.22
6	PRHLT	90.80
7	NLEL	62.78

Table 6: Closed submission results for test set B.

4.3 Open Submission

Three teams participated in the open submission track: NRC, NLEL, and OSEVAL. Their results are shown in Table 7. Unlike DSL 2014 (see Table 1), two of these teams, NRC and NLEL, managed to achieve better accuracy in the open submission than in the closed one on test set A.[8]

[7] More detailed evaluation results can be found at `https://github.com/Simdiva/DSL-Task/blob/master/DSL2015-results.md`

[8] OSEVAL did not participate in the closed submission.

Rank	Team	Accuracy
1	NRC	95.65
2	NLEL	91.84
3	OSEVAL	76.17

Table 7: Open submission results for test set A.

This could be related to the availability of DSLCC v1.0 as an obvious additional resource. The NRC system description paper indeed confirms that they used DSLCC v1.0 (Goutte and Léger, 2015), and points out that this yielded 10% error reduction and 0.4% absolute boost in accuracy. In contrast, teams that submitted open submissions to the 2014 edition did not have access to such a well-matching additional resource.

The open submission results for test set B are shown in Table 8: we can see once again improved performance for NLEL and NRC.[9]

Rank	Team	Accuracy
1	NRC	93.41
2	NLEL	89.56
3	OSEVAL	75.30

Table 8: Open submission results for test set B.

4.4 Results per Language

Not all language pairs and groups of languages are equally difficult to distinguish from the rest. We wanted to have a closer look at this, and thus we plotted for each language the mean accuracy across all submissions and the interquartal range, excluding outliers: accuracy results for test sets A and B in the closed submission track are shown in Figures 1 and 2, respectively.

We can see that, on test set A, systems performed very well when discriminating between the languages in the following pairs: Bulgarian–Macedonian, Czech–Slovak, and Indonesian–Malay. On test set B, distinguishing between Indonesian and Malay was difficult, maybe because there were many country-specific named entities in Indonesian and Malay texts, which were helping to discriminate between them on test set A. Overall, the most challenging groups are Bosnian–Croatian–Serbian, as well as the Spanish and the Portuguese varieties, which corroborates the findings of the first edition of the DSL shared task.

5 Approaches

The participants used a variety of classifiers and features, which, in our opinion, confirms the DSL shared task as a very fruitful scientific endeavor for both organizers and participants.

The best system in the closed submission was that of the MAC team (Malmasi and Dras, 2015b). They used an ensemble of SVM classifiers, and features such as character n-grams (n=1,2,...,6) and word unigrams and bigrams.

The NRC team (Goutte and Léger, 2015) included members of the NRC-CNRC team, which won the DSL closed submission track in 2014. Both in 2014 and now, they used two-stage classification, which first predicts the language group, and then chooses between languages or varieties within this group. The team achieved very strong results this year, ranking second in the closed submission on test set A, third on test set B, and first in the open submission on both test sets A and B. Two other participants used two-stage classification: NLEL (Fabra-Boluda et al., 2015) and BRUniBP (Ács et al., 2015).

The MMS team experimented with three approaches (Zampieri et al., 2015), and their best run combined TF.IDF weighting and an SVM classifier, which was previously successfully applied to native language identification (Gebre et al., 2013).

The SUKI team (Jauhiainen et al., 2015a) used *token-based backoff*, which was previously applied to general-purpose language identification (Jauhiainen et al., 2015b).

The BOBICEV team applied *prediction by partial matching*, which had not been used for this task before (Bobicev, 2015).

Finally, the PRHLT team (Franco-Salvador et al., 2015) used word and sentence vectors, which is to our knowledge the first attempt to apply them to discriminating between similar languages.

6 Conclusion

The second edition of the DSL shared task, with its focus on similar languages, continues to fill an important gap in language identification research. It allows researchers to experiment with different algorithms and methods and to evaluate their systems for discriminating between related languages and language varieties. Compared to the first edition, this year we observed an increase in team participation, which shows the continuous interest of the research community in this task.

[9] Note, however, that NLEL reported having a bug, which is an alternative explanation for the low performance of their closed submission runs.

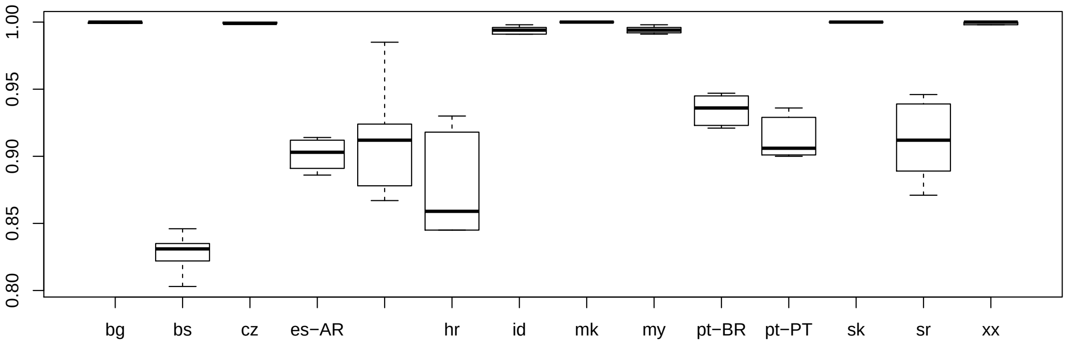

Figure 1: Accuracy per language: closed submission, test set A.

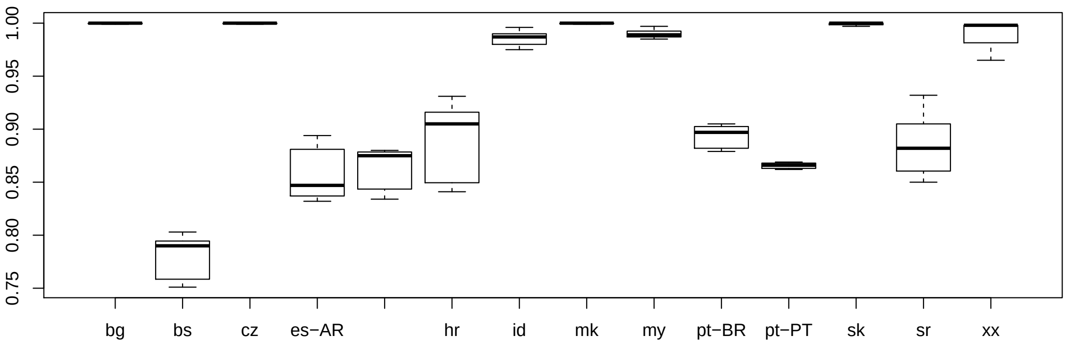

Figure 2: Accuracy per language: closed submission, test set B.

In total, 24 teams registered to participate, and 10 made submissions. The best-performing system in the closed submission track was that of MAC (Malmasi and Dras, 2015b), and it achieved 95.54% accuracy on test set A and 94.01% on test set B, using an ensemble of SVM classifiers. The winner in the open submission track NRC (Goutte and Léger, 2015) achieved 95.65% accuracy on test set A, and 93.41% on test set B, using two-stage classification.

Unlike the 2014 edition, in 2015 we had the *Others* category with languages not seen on training. Moreover, we had a second test set, where named entities were replaced by placeholders.

Comparing the results for the two test sets, (*i*) the original vs. (*ii*) the one with placeholders, has shown that the accuracy on the latter dropped by about 2% absolute for all teams. However, the impact of substituting named entities was not as great as we had imagined, especially for language groups for which the accuracy was already close or equal to 100% (except for Indonesian–Malay). This suggests that closely-related languages and language varieties have distinctive properties that classifiers are able to recognize and learn.

For a possible third edition of the DSL Shared Task, we would like to explore the possibility to include dialects in the dataset. The case of Arabic is particularly interesting, and has already attracted research attention (Sadat et al., 2014). Unfortunately, Arabic dialects do not have official status and thus are not common in journalistic texts; thus, we would need to compile a heterogeneous dataset including other genres as well.

Another interesting aspect, which we did not study explictly in the first two editions of the DSL Shared Task (even though the instances in v1.0 and v2.0 of DSLCC did have different length distributions), but which we would like to explore in the future, is the influence of text length on the classification performance. See (Malmasi et al., 2015) for a relevant discussion.

Acknowledgements

We would like to thank all participants in the DSL Shared Task for their valuable suggestions and comments. We further thank the LT4VarDial Program Committee for thoroughly reviewing the system papers and the shared task report.

References

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Timothy Baldwin and Marco Lui. 2010. Multilingual language identification: ALTW 2010 shared task data. In *Proceedings of Australasian Language Technology Association Workshop*, ALTA '10, pages 4–7, Melbourne, Australia.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig corpora collection-monolingual corpora of standard size. In *Proceedings of Corpus Linguistics*, Birmingham, UK.

Victoria Bobicev. 2015. Discriminating between similar languages using PPM. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 627–632, Doha, Qatar.

Raül Fabra-Boluda, Francisco Rangel, and Paolo Rosso. 2015. NLEL UPV autoritas participation at Discrimination between Similar Languages (DSL) 2015 shared task. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Marc Franco-Salvador, Paolo Rosso, and Francisco Rangel. 2015. Distributed representations of words and documents for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with TF-IDF weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, BEA8, pages 216–223, Atlanta, Georgia, USA.

Cyril Goutte and Serge Léger. 2015. Experiments in discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, VarDial '14, pages 139–145, 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? In *Actes du sixième Défi Fouille de Textes*, DEFT '10, Montreal, Canada.

Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of 22nd Pacific Asia Conference on Language, Information and Computation*, PACLIC '08, pages 404–410, Cebu City, Philippines.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2015a. Discriminating similar languages with token-based backoff. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2015b. Language set identification in noisy synthetic multilingual documents. In *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLING '15, pages 633–643, Cairo, Egypt.

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 '14, pages 146–154, 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*, IJCNLP '11, pages 553–561, Chiang Mai, Thailand.

Marco Lui and Timothy Baldwin. 2012. Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 25–30, Jeju Island, Korea.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of Australasian Language Technology Workshop*, ALTA '13, pages 5–15, Brisbane, Australia.

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 '14', pages 129–138, Dublin, Ireland.

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

Shervin Malmasi and Mark Dras. 2015b. Language identification using classifier ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, Hissar, Bulgaria.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics*, PACLING '15, pages 209–217, Bali, Indonesia.

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 '14, pages 120–128, 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 Dialect*, VarDial '14, pages 155–160, 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.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media*, SocialNLP '14, pages 22–27, Dublin, Ireland.

Alberto Simões, José João Almeida, and Simon D Byers. 2014. Language identification: a neural network approach. In *Proceedings of the 3rd Symposium on Languages, Applications and Technologies*, SLATE '14, pages 252–265, Dagstuhl, Germany.

Noah A. Smith, Claire Cardie, Anne L. Washington, and John D. Wilkerson. 2014. Overview of the 2014 NLP unshared task in PoliInformatics. In *Proceedings of the Workshop on Language Technologies and Computational Social Science*, pages 5–7, Baltimore, Maryland, USA.

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.

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 '14, pages 6–10, Reykjavik, Iceland.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING '12, pages 2619–2634, Mumbai, India.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC '12, pages 2214–2218, Istanbul, Turkey.

Francis Tyers and Murat Serdar Alperen. 2010. South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, Valetta, Malta.

Fei Xia, Carrie Lewis, and William D Lewis. 2010. The problems of language identification within hugely multilingual data sets. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC '10, pages 2790–2797, Valetta, Malta.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, Nashville, Tennessee, USA.

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

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, VarDial '14, pages 58–67, Dublin, Ireland.

Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. 2015. Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, 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 '14, pages 1–11, Girona, Spain.