Findings of the VarDial Evaluation Campaign 2017

Zampieri, Marcos; Malmasi, Shervin; Ljubešić, Nikola; Nakov, Preslav; Ali, Ahmed; Tiedemann, Jörg; Scherrer, Yves; Aepli, Noëmi

Findings of the VarDial Evaluation Campaign 2017

Marcos Zampieri[1], Shervin Malmasi[2,3], Nikola Ljubešić[4,5], Preslav Nakov[6]
Ahmed Ali[6], Jörg Tiedemann[7], Yves Scherrer[8], Noëmi Aepli[9]

[1]University of Cologne, Germany, [2]Harvard Medical School, USA
[3]Macquarie University, Australia, [4]University of Zagreb, Croatia
[5]Jožef Stefan Institute, Slovenia, [6]Qatar Computing Research Institute, HBKU, Qatar
[7]University of Helsinki, Finland, [8]University of Geneva, Switzerland
[9]University of Zurich, Switzerland

Abstract

We present the results of the VarDial Evaluation Campaign on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects, which we organized as part of the fourth edition of the VarDial workshop at EACL'2017. This year, we included four shared tasks: Discriminating between Similar Languages (DSL), Arabic Dialect Identification (ADI), German Dialect Identification (GDI), and Cross-lingual Dependency Parsing (CLP). A total of 19 teams submitted runs across the four tasks, and 15 of them wrote system description papers.

1 Introduction

The VarDial Evaluation Campaign targets Natural Language Processing (NLP) for similar languages, varieties and dialects, and it was organized within the scope of the VarDial'2017 workshop. The campaign is an evolution of the DSL shared tasks, which were organized as part of the previous editions of the VarDial workshop (Zampieri et al., 2014; Zampieri et al., 2015b; Malmasi et al., 2016), and which have focused on the discrimination of similar languages and language varieties as well as on dialect identification.

Since the first DSL challenge, we have observed a substantial increase in the interest from the community. The 2016 edition of the DSL task, which included a sub-task on Arabic Dialect Identification, attracted a notably larger number of participants compared to the previous two editions. Thus, we decided to further extend the scope of the shared task, turning it into a more comprehensive evaluation campaign with several independent shared tasks, which included but were not limited to dialect and similar language identification.

1.1 Shared Tasks

The VarDial Evaluation Campaign 2017 included four tasks:

Discriminating between Similar Languages (DSL): This was the fourth iteration of the multilingual similar language and language variety identification task. The goal was to recognize the language of short excerpts of texts extracted from newspapers. This included several similar languages and language varieties: Bosnian, Croatian, and Serbian; Malay and Indonesian; Persian and Dari; Canadian and Hexagonal French; Brazilian and European Portuguese; Argentinian, Peninsular, and Peruvian Spanish.

Arabic Dialect Identification (ADI): This was the second iteration of the ADI task, which was organized as a sub-task of the DSL task in 2016 (Malmasi et al., 2016). The goal was to recognize the dialect of speech transcripts along with acoustic features. The following Arabic dialects were included: Egyptian, Gulf, Levantine, North-African, and Modern Standard Arabic (MSA).

German Dialect Identification (GDI): This task included Swiss German dialects from four areas: Basel, Bern, Lucerne, and Zurich. We provided manually annotated speech transcripts for all dialect areas; unlike ADI, we provided no acoustic data for this task.

Cross-lingual Dependency Parsing (CLP): The task is to parse some *target language* (TL) without annotated training data for that language but given annotated data for a closely related-language(s), called *source language* (SL). We included the following language pairs: Croatian (TL) – Slovenian (SL), Slovak (TL) – Czech (SL), Norwegian (TL) – Danish, and Norwegian (TL) – Swedish (SL). Note that the latter two pairs include a triple of related languages.

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 1–15,
Valencia, Spain, April 3, 2017.

Team	DSL	ADI	GDI	CLP	System Description Paper
ahaqst		✓	✓		(Hanani et al., 2017)
bayesline	✓				–
CECL	✓		✓		(Bestgen, 2017)
cic_ualg	✓				(Gómez-Adorno et al., 2017)
Citius_Ixa_Imaxin	✓		✓		(Gamallo et al., 2017)
CLUZH			✓		(Clematide and Makarov, 2017)
CUNI				✓	(Rosa et al., 2017)
deepCybErNet	✓	✓	✓		–
gauge	✓				–
Helsinki-CLP				✓	(Tiedemann, 2017)
MAZA (ADI)		✓			(Malmasi and Zampieri, 2017a)
MAZA (GDI)			✓		(Malmasi and Zampieri, 2017b)
mm_lct	✓				(Medvedeva et al., 2017)
qcri_mit		✓	✓		–
SUKI	✓				(Jauhiainen et al., 2017)
timeflow	✓				(Criscuolo and Aluisio, 2017)
tubasfs	✓	✓	✓	✓	(Çöltekin and Rama, 2017)
unibuckernel		✓	✓		(Ionescu and Butnaru, 2017)
XAC_Bayesline	✓		✓		(Barbaresi, 2017)
Total	**11**	**6**	**10**	**3**	**15**

Table 1: The teams that participated in the VarDial'2017 Evaluation Campaign.

1.2 Participating Teams

The VarDial Evaluation Campaign received a positive response from the research community: a total of 26 teams enrolled to participate, 19 teams eventually submitted systems, and 15 of them wrote system description papers. Table 1 lists the participating teams and the shared tasks they took part in.[1] We can see that each task received multiple submissions, ranging from 3 for CLP to 11 for DSL. Below we describe the individual tasks.

2 Discriminating between Similar Languages (DSL)

Discriminating between similar languages is one of the main challenges faced by language identification systems. Since 2014 the DSL shared task has been organized every year providing scholars and developers with an opportunity to evaluate language identification methods using a standard dataset and evaluation methodology. Albeit related to other shared tasks such as the 2014 *Tweet-LID* challenge (Zubiaga et al., 2014) and the 2016 shared task on Geolocation Prediction (Han et al., 2016), the DSL shared task continues to be the only shared task focusing on the discrimination between similar languages and language varieties.

The fourth edition of the DSL shared task was motivated by the success of the previous editions and by the growing interest of the research community in the identification of dialects and similar languages, as evidenced by recent publications (Xu et al., 2016; Radford and Gallé, 2016; Castro et al., 2016). We also saw the number of system submissions to the DSL challenge grow from 8 in 2014 to 10 in 2015 and then to 17 in 2016.[2]

The 2015 and the 2016 editions of the DSL task focused on one under-explored aspect of the task in order to keep it interesting and challenging.

In 2015 (Zampieri et al., 2015b), we investigated the extent to which named entities influenced system performance. Obviously, newspapers from Brazil mention *Rio de Janeiro* more often than those in Portugal do, and Argentinian newspapers talk more about *Buenos Aires* than those in Spain. In order to investigate this aspect, in 2015 we provided participants with two test sets, one containing the original unmodified texts (test set A) and another one containing texts with capitalized named entities substituted by placeholders (test set B). Eventually, we observed that the impact of named entities was not as sizable as we had anticipated.

[1] The MAZA team submitted two separate papers: one for each task they participated in.

[2] This number does not include the submissions to the Arabic Dialect Identification subtask of DSL in 2016.

At DSL 2015, the four best systems, MAC (Malmasi and Dras, 2015b), MMS (Zampieri et al., 2015a), NRC (Goutte and Léger, 2015), and SUKI (Jauhiainen et al., 2015) performed similarly on test set B compared to test set A: in the closed training setting, where the systems were trained only using the training data provided by the DSL organizers, their accuracy dropped from 95.54 to 94.01, from 95.24 to 92.78, from 95.24 to 93.01, and from 94.67 to 93.02, respectively.[3]

Finally, inspired by recent work on language identification of user-generated content (Ljubešić and Kranjčić, 2015; Abainia et al., 2016), in the DSL 2016 task (Malmasi et al., 2016), we looked at how systems perform on discriminating between similar languages and language varieties across different domains, an aspect highlighted by Lui and Cook (2013) and Lui (2014). For this purpose, we provided an out-of-domain test set containing manually annotated microblog posts written in Bosnian, Croatian, Serbian, Brazilian and European Portuguese.

2.1 Task Setup

We applied the methodology of Tan et al. (2014) in order to compile version 4.0 of the DSL Corpus Collection (DSLCC), which contains short excerpts of journalistic texts; we describe the corpus in detail in Section 2.2 below.

We first released the training and the development datasets, in which all instances were labeled with the correct language or language variety. One month later, the participants received an unlabeled test set, which they had to annotate with their system's prediction. The participating teams were allowed to use the DSLCC v4.0 corpus or any other dataset, and we had two types of training conditions.

- **Closed Training:** using only the corpora provided by the organizers (DSLCC v4.0);
- **Open Training:** using any additional data including previous versions of the DSLCC corpus.

For each kind of training, we allowed a maximum of three runs per team, i.e., six in total.

[3]For a comprehensive evaluation of the 2014 and 2015 editions of the DSL shared task see (Goutte et al., 2016).

2.2 Dataset

The DSLCC v4.0[4] contains 22,000 short excerpts of news texts for each language or language variety divided into 20,000 texts for training (18,000 texts) and development (2,000 texts), and 2,000 texts for testing. It contains a total of 8.6 million tokens for training and over half a million tokens for testing. The fourteen languages included in the v4.0 grouped by similarity are Bosnian, Croatian, and Serbian; Malay and Indonesian; Persian and Dari; Canadian and Hexagonal French; Brazilian and European Portuguese; Argentinian, Peninsular, and Peruvian Spanish. In Table 2, we present the number of instances and the total number of documents and tokens we released for each language or language variety.

As indicated in Table 2, some languages were available in all previous versions of the DSLCC corpus (e.g., Bosnian, Croatian, and Serbian) or only in some of them (e.g., Canadian and Hexagonal French). As v4.0 is comparable to the previous versions of the DSLCC, this provided teams with more training data to use in the open training track.

Note that Peruvian Spanish, Persian, and Dari appear for the first time in the DSL task. However, they were previously included in language identification experiments: Peruvian Spanish was used in four-way classification together with texts from Argentina, Mexico, and Spain, for which an F1 of 0.876 was reported (Zampieri et al.,), and there were previous experiments in discriminating between Persian and Dari, which achieved 0.96 accuracy (Malmasi and Dras, 2015a).

2.3 Participants and Approaches

Twenty teams enrolled to participate in this edition of the DSL shared task and eleven of them submitted results. This represents a slight decrease in participation compared to the 2016 edition, which followed an uphill trend in participation since the first DSL organized in 2014. In our opinion, this slight decrease in participation does not represent less interest of the scientific community in the topic. Discriminating between similar languages and language varieties continues to be a vibrant research topic and the interest of the community is confirmed by the recent aforementioned publications (Xu et al., 2016; Radford and Gallé, 2016).

[4]All versions of the DSLCC dataset are available at `http://ttg.uni-saarland.de/resources/DSLCC`

		Train & Dev.		Test		Previous DSLCC		
Language/Variety	**Class**	**Instances**	**Tokens**	**Instances**	**Tokens**	**v1.0**	**v2.0/2.1**	**v3.0**
Bosnian	bs	20,000	716,537	1,000	35,756	✓	✓	✓
Croatian	hr	20,000	845,639	1,000	42,774	✓	✓	✓
Serbian	sr	20,000	777,363	1,000	39,003	✓	✓	✓
Indonesian	id	20,000	800,639	1,000	39,954	✓	✓	✓
Malay	my	20,000	591,246	1,000	29,028	✓	✓	✓
Brazilian Portuguese	pt-BR	20,000	907,657	1,000	45,715	✓	✓	✓
European Portuguese	pt-PT	20,000	832,664	1,000	41,689	✓	✓	✓
Argentine Spanish	es-AR	20,000	939,425	1,000	42,392	✓	✓	✓
Castilian Spanish	es-ES	20,000	1,000,235	1,000	50,134	✓	✓	✓
Peruvian Spanish	es-PE	20,000	569,587	1,000	28,097			
Canadian French	fr-CA	20,000	712,467	1,000	36,121			✓
Hexagonal French	fr-FR	20,000	871,026	1,000	44,076			✓
Persian	fa-IR	20,000	824,640	1,000	41,900			
Dari	fa-AF	20,000	601,025	1,000	30,121			
Total		280,000	8,639,459	14,000	546,790			

Table 2: DSLCC v4.0: the languages included in the corpus grouped by similarity.

The slight decrease in participation is largely due to bad timing. Because of EACL-related deadlines, DSL 2017 was organized only a few months after the 2016 edition had finished, and the training data was released between Christmas and New Year's Eve. Moreover, this year the DSL was not a standalone task,[5] and it was part of a larger evaluation campaign. This has resulted in participants splitting between the four tasks we were running as part of the VarDial Evaluation Campaign. Yet, the DSL task attracted the highest number of participants, both new and returning.

We find a variety of computational approaches and features used by the participating systems. Below, we present a brief overview of each submission, ordered by the weighted F1 score. The interested reader can find more information about an individual system in the respective system description paper, which is referred to in the last column of Table 1.

- **CECL:** The system uses a two-step approach as in (Goutte et al., 2014). The first step identifies the language group using an SVM classifier with a linear kernel trained on character n-grams (1-4) that occur at least 100 times in the dataset weighted by Okapi BM25 (Robertson et al., 1995). The second step discriminates between each language within the group using a set of SVM classifiers trained on a variety of features such as character n-grams of various orders, global statistics such as proportion of capitalized letters, punctuation marks, and spaces, and finally POS tags modeled as n-grams (1-5) for French, Portuguese, and Spanish obtained by annotating the corpus using TreeTagger (Schmid, 1994).

- **mm_lct:** This team submitted three runs. Run 1 (their best) used seven SVM classifiers in two steps. First, one SVM classifier finds the language group, and then six individual SVM classifiers distinguish between the languages in each group. Run 2 used a linear-kernel SVM trained using word n-grams (1–2) and character n-grams (up to 6). Run 3 used a recurrent neural network (RNN).

- **XAC_Bayesline:** This system is a refined version of the Bayesline system (Tan et al., 2014), which was based on character n-grams and a Na"ive Bayes classifier. The system followed the work of the system submitted to the DSL 2016 by Barbaresi (2016).

- **tubasfs:** Following the success of tubasfs at DSL 2016 (Çöltekin and Rama, 2016), which was ranked first in the closed training track, this year's tubasfs submission used a linear SVM classifier. The system used both characters and words as features, and carefully optimized hyperparameters: n-gram size and margin/regularization parameter for SVM.

[5] In 2016 ADI and DSL were organized under the name *DSL shared task*, and ADI was run as a sub-task.

- **gauge:** This team submitted a total of three runs. Run 1 used an SVM classifier with character n-grams (2–6), run 2 (their best run) used logistic regression trained using character n-grams (1–6), and run 3 used hard voting of three systems: SVM, Logistic Regression, and Na"ive Bayes and character n-grams (2–6) as features.

- **cic_ualg:** This team submitted three runs. Runs 1 and 2 first predict the language group, and then discriminate between the languages within that group. The first step uses an SVM classifier with a combination of character 3–5-grams, typed character 3-grams, applying the character n-gram categories introduced by Sapkota et al. (2015), and word unigrams using TF-weighting. The second step uses the same features and different classifiers: SVMs + Multinominal Naïve Bayes (MNB) in run 1, and MNB in run 2 (which works best). Run 3 uses a single MNB classifier to discriminate between all fourteen languages.

- **SUKI:** This team's submission was based on the token-based backoff method used in SUKI's DSL submission in 2015 (Jauhiainen et al., 2015) and in 2016 (Jauhiainen et al., 2016). Run 1 used character 1–8-grams, and run 2 (their best) used loglike mapping (Brown, 2014) instead of relative frequencies, together with character 1–7-grams.

- **timeflow:** This system used a two-step classifier, as introduced by Goutte et al. (2014); a similar approach was used by some other teams. First, they used a Na"ive Bayes classifier trained on character n-grams to detect the language group. Then, they distinguished the language or language variety within the detected group using Convolutional Neural Networks (CNNs) with learned word embeddings and Multi-Layer Perceptron (MLP) with TF.IDF vectors.

- **Citius_Ixa_Imaxin:** This team was the only one to participate in both the open and the closed tracks. Their system was based on language model perplexity. The best performance in the closed training condition was obtained in run 1, which applied a voting scheme over 1–3 word n-gram and 5–7 characters n-grams.

- **bayesline:** This team participated with a Multinomial Naïve Bayes (MNB) classifier similar to that of Tan et al. (2014), with no special parameter tuning, as this system was initially intended to serve as an intelligent baseline for the task (but now it has matured into a competitive system). In their best-performing run 1, they relied primarily on character 4-grams as features. The feature sets they used were selected by a search strategy as proposed in (Scarton et al., 2015).

- **deepCybErNet:** This team approached the task using a neural network based on Long Short-Term Memory (LSTM). Neural networks have been successfully applied to several NLP tasks in recent years, but the results of the deepCybErNet team in the DSL and the GDI tasks in 2017, as well as in DSL 2016 (Malmasi et al., 2016), suggest that using neural networks is of limited use in our limited training data scenario: neural networks have many parameters to optimize, which takes a lot of training data, much more than what we provide here.

2.4 Results

Only one team, *Citius_Ixa_Imaxin*, submitted results to the open training track, achieving 0.9 accuracy. As there were no other submissions to compare against, in this section we report and discuss the results obtained by participants in the closed training track only.

Table 3 presents the best results obtained by the participating teams. We rank them based on their weighted F1 score (weighted by the number of examples in each class).

Rank	Team	F1 (weighted)
1	CECL	0.927
2	mm_lct	0.925
3	XAC_Bayesline	0.925
4	tubasfs	0.925
5	gauge	0.916
6	cic_ualg	0.915
7	SUKI	0.910
8	timeflow	0.907
9	Citius_Ixa_Imaxin	0.902
10	bayesline	0.889
11	deepCybErNet	0.202

Table 3: DSL task: closed submission results.

The *CECL* team achieved best performance: F1=0.927. It is followed by three teams, all tied with an F1 score of 0.925: namely *mm_lct*, *XAC_Bayesline*, and *tubasfs*.

The system description paper of *CECL* (Bestgen, 2017) provides some interesting insights about the DSL task. First, they found out that BM25 weighting, which was previously applied to native language identification (NLI) (Wang et al., 2016), worked better than using TF.IDF. They further highlighted the similarity between similar language identification and NLI as evidenced by a number of entries in the DSL task that are adaptations of systems used for NLI (Goutte et al., 2013; Gebre et al., 2013; Jarvis et al., 2013).

We observe that the variation in performance among the top ten teams is less than four percentage points. The team ranked last (eleventh) approached the task using LSTM and achieved an F1 score of 0.202. Unfortunately, they did not submit a system description paper, and thus we do not have much detail about their system. However, in the DSL 2016 task (Malmasi et al., 2016), neural network-based approaches already proved not to be very competitive for the task. See (Medvedeva et al., 2017) for a comparison between the performance of an SVM and an RNN approach for the DSL task.

2.5 Summary

The fourth edition of the DSL shared task allowed us once again to compare a variety of approaches for the task of discriminating between similar languages and language varieties using the same dataset: DSLCC v4.0. Even though previous versions of the DSLCC were available for use in an open track condition, all teams with the exception of *Citius_Ixa_Imaxin* chose to compete in the closed training track only.

The participants took advantage of the experience acquired in the previous editions of the DSL task, and in absolute terms achieved the highest scores among all four editions of the DSL challenge. *CECL* achieved 0.927 F1-score and *mm_lct*, *XAC_Bayesline*, and *tubasfs* achieved 0.925.

For the reasons discussed in Section 2.3, the participation in the DSL 2017 was slightly lower than in the 2016 edition, but it was still higher than in 2014 and 2015.

3 Arabic Dialect Identification (ADI)

The ADI task was introduced in 2016 (Malmasi et al., 2016), where it was run as a subtask of the DSL task. Unlike the DSL task, which is about text, the ADI task is based on speech transcripts, as Arabic dialects are mostly used in conversation. The ADI task asks to discriminate at the utterance level between five Arabic varieties, namely Modern Standard Arabic (MSA) and four Arabic dialects: Egyptian (EGY), Gulf (GLF), Levantine (LAV), and North African (NOR).

This year's edition of the task was motivated by the success of the 2016 edition and by the growing interest in dialectal Arabic in general. In 2016, we provided task participants with input speech transcripts generated using Arabic Large Vocabulary Speech Recognition (LVCSR) following the approach in (Ali et al., 2014a), from which we further extracted and provided lexical features. This year, we added a multi-model aspect to the task by further providing acoustic features.

3.1 Dataset

As we said above, this year we used both speech transcripts and acoustic features. The speech transcription was generated by a multi-dialect LVCSR system trained on 1,200+ speech hours for acoustic modeling and on 110+ million words for language modeling; more detail about the system, which is the winning system of the Arabic Multi-Genre Broadcast (MGB-2) challenge, can be found in (Khurana and Ali, 2016).

For the acoustic features, we released a 400-dimensional i-vector for each utterance. We extracted these i-vectors using Bottle Neck Features (BNF) trained on 60 hours of speech data; see (Ali et al., 2016) for detail.

The data for the ADI task comes from a multi-dialectal speech corpus created from high-quality broadcast, debate and discussion programs from Al Jazeera, and as such contains a combination of spontaneous and scripted speech (Wray and Ali, 2015). We collected the training dataset from the Broadcast News domain in four Arabic dialects (EGY, LAV, GLF, and NOR) as well as in MSA. The audio recordings were carried out at 16Khz. The recordings were then segmented in order to avoid speaker overlap, also removing any non-speech parts such as music and background noise; more detail about the training data can be found in (Bahari et al., 2014).

Dialect	Dialect	Training			Development			Testing		
		Ex.	Dur.	Words	Ex.	Dur.	Words	Ex.	Dur.	Words
Egyptian	EGY	3,093	12.4	76	298	2	11.0	302	2.0	11.6
Gulf	GLF	2,744	10.0	56	264	2	11.9	250	2.1	12.3
Levantine	LAV	2,851	10.3	53	330	2	10.3	334	2.0	10.9
MSA	MSA	2183	10.4	69	281	2	13.4	262	1.9	13.0
North African	NOR	2,954	10.5	38	351	2	9.9	344	2.1	10.3
Total		13,825	53.6	292	1524	10	56.5	1492	10.1	58.1

Table 4: The ADI data: examples (Ex.) in utterances, duration (Dur.) in hours, and words in 1000s.

Although the test and the development datasets came from the same broadcast domain, the recording setup was different from the training data. We downloaded the test and the development data directly from the high-quality video server for Al Jazeera (brightcove) over a period between July 2104 and January 2015, as part of QCRI's Advanced Transcription Service (QATS) (Ali et al., 2014b). In addition to the lexical and the acoustic features, we also released the audio files.[6] Table 4 shows some statistics about the ADI training, development and testing datasets.

3.2 Participants and Approaches

We received six submissions for the ADI task, all for the closed training condition. The teams below are sorted according to their performance on the test dataset.

- **unibuckernel:** This team submitted two runs. Run 1 was a Kernel Ridge Regression (KRR) classifier trained on the sum of a blended presence bits kernel based on 3–5-grams, a blended intersection kernel based on 3–7-grams, a kernel based on Local Rank Distance (LRD) with n-grams of 3 to 7 characters, and a quadratic RBF kernel based on i-vectors. This setup achieved an F1 of 0.642 on the development set, and 0.763 on the test set. Run 2 was a Kernel Discriminant Analysis (KDA) classifier trained on the sum of a blended presence bits kernel using 3–5-grams, a blended intersection kernel based on 3–7-grams, a kernel based on LRD with 3 to 7 characters, and a quadratic RBF kernel based on i-vectors. This setup achieved an F1 of 0.75 on the test set. More detail can be found in (Ionescu and Butnaru, 2017).

- **MAZA:** This team submitted three runs. Run 1 was a voting ensemble (F1=0.72), run 2 was a mean probability ensemble (F1=0.67), and run 3 was a meta classifier (F1=0.61). They used character 1–8-grams, word unigrams, and i-vectors. More detail about the system can be found in (Malmasi and Zampieri, 2017a).

- **tubasfs:** This team submitted two runs. Run 1 used a linear SVM with words and i-vectors, achieving an F1 of 0.70. Run 2 only used word features, which yielded an F1 of 0.57. More detail about the system can be found in (Çöltekin and Rama, 2017).

- **ahaqst:** This team submitted three runs. Run 1 used a focal multiclass model to combine the outputs of a word-based SVM multiclass model, and an i-vector-based SVM multiclass model, achieving an F1 of 0.63. Run 2 combined Na"ive Bayes with multinomial distribution, SVM with a Radial Basis Function (RBF) kernel, logistic regression, and Random Forests with 300 trees, achieving an F1 of 0.31. Run 3 combined five systems, which used WAV files only for recognizing Arabic dialects, i-Vectors plus Gaussian Mixture Model-Universal Background Model (GMM-UBM) plus phonotactic plus GMM tokenization (256 bigrams and 20,148 unigrams), achieving an F1 of 0.59. More detail about their system can be found in (Hanani et al., 2017).

- **qcri_mit:** This team submitted three runs. Run 1 combined (*i*) normalized scores from an SVM model trained on Latent Dirichlet Allocation (LDA) i-vectors (down to a 4-dimensional vector) with (*ii*) an SVM classifier trained on character 1–4-grams, achieving an F1 score of 0.616. Run 2 combined

[6] https://github.com/Qatar-Computing-Research-Institute/dialectID/tree/master/data

(*i*) an SVM using LDA with Within-Class Covariance Normalization (WCCN) i-vector with (*ii*) an SVM trained on count-based bag of character 2–6-grams, achieving an F1 of 0.615. Run 3 combined (*i*) an SVM model using LDA with WCCN i-vector (as in Run 2) with (*ii*) an SVM model trained on count bag of characters 2–4-grams, which yielded an F1 of 0.612.

- **deepCybErNet:** This team submitted two runs. Run 1 adopted a Bi-LSTM architecture using the lexical features, and achieved an F1 score of 0.208, while run 2 used the i-vector features and achieved an F1 of 0.574.

3.3 Results

Table 5 shows the evaluation results for the ADI task. Note that those participants who had used the development data for training their models obtained substantial gains, e.g., the winning system *unibuckernel* achieved an F1 of 0.763. However, this same system would have scored only 0.611, had they trained on the training data only. We attribute this to both the development and the testing data coming from the recording setup, and that is why using the i-vectors particularity has helped to model the channel, not only the dialect.

Rank	Team	F1 (weighted)
1	unibuckernel	0.763
2	MAZA	0.717
3	tubasfs	0.697
4	ahaqst	0.628
5	qcrimit	0.616
6	deepCybErNet	0.574

Table 5: ADI task: closed submission results.

3.4 Summary

This year's ADI task was very successful, as for the first time in VarDial the participants were provided with acoustic features. Indeed, as we have seen above, the i-vectors were widely used by the participating teams. Most participants took advantage of the fact that the development data came from the same recording setup as the testing data, which has boosted their results. Moreover, one team used the raw audio files. In the future, we plan another iteration of the task, where we would add phonotactic features and phoneme duration.

4 German Dialect Identification (GDI)

This year, we introduced a new dialectal area, which focused on German dialects of Switzerland. Indeed, the German-speaking part of Switzerland is characterized by the widespread use of dialects in everyday communication, and by a large number of different dialects and dialectal areas.

There have been two major approaches to Swiss German dialect identification in the literature. The *corpus-based approach* predicts the dialect of any text fragment extracted from a corpus (Scherrer and Rambow, 2010; Hollenstein and Aepli, 2015). The *dialectological approach* tries to identify a small set of distinguishing dialectal features, which are then elicited interactively from the user in order to identify his or her dialect (Leemann et al., 2016). In this task, we adopt a corpus-based approach, and we develop a new dataset for this.

4.1 Dataset

We extracted the training and the test datasets from the ArchiMob corpus of Spoken Swiss German (Samardžić et al., 2016). The current release of the corpus contains transcriptions of 34 oral history interviews with informants speaking different Swiss German dialects.

Each interview was transcribed by one of four transcribers, using the writing system "Schwyzertütschi Dialäktschrift" proposed by Dieth (1986). The transcription is expected to show the phonetic properties of the variety, but in a way that is legible for everybody who is familiar with the standard German orthography. Although its objective is to keep track of the pronunciation, Dieth's transcription method is orthographic and partially adapted to the spelling habits in standard German. Therefore, it does not provide the same precision and explicitness as phonetic transcription methods do. Moreover, the transcription choices are dependent on the dialect, the accentuation of the syllables and – to a substantial degree – also the dialectal background of the transcriber. Also, the practice of using Dieth's system changed over time, so that some transcribers (e.g., transcriber P in Table 6) made more distinctions concerning the openness of vowels than others. The transcriptions exclusively used lowercase. Note that Dieth's system is hardly known by laymen, so that Swiss German data extracted from social media would look fairly different from our transcripts.

Dialect	Doc.	Utter.	Trans.	Dist.
BE	1142	794	P	<5
	1170	872	P	45
	1215	2,223	M	13
	1121*	906	M	<5
BS	1044	952	A	<5
	1073	1,407	P	23
	1075	1,052	P	<5
	1263*	939	A	<5
LU	1007	815	P	11
	1195	1,070	P	13
	1261	1,329	P	<5
	1008*	916	A	5
ZH	1082	842	M	<5
	1087	933	M	<5
	1143	759	P	6
	1244	728	M	19
	1270	702	P	6
	1225*	877	M	<5

Table 6: ArchiMob interviews used for the GDI task. *Doc.* = document identifier (starred identifiers refer to the test set), *Utter.* = number of utterances included in the GDI dataset, *Trans.* = identifier of the transcriber, *Dist.* = distance (in kilometers) from the core city of the dialect area.

We have been able to identify four dialectal areas for which sufficient amounts of data were available and which were known to be distinct enough. The selected dialect areas correspond to four large agglomerations in the German-speaking part of Switzerland: Zurich (ZH), Basel (BS), Bern (BE), and Lucerne (LU).

The training set contains utterances from at least 3 interviews per dialect, and the test set contains utterances from another interview (see Table 6). The data were sampled such that at least one of the training interviews was transcribed by the same transcriber as the corresponding test interview, except for LU. For LU and BS, we included additional transcripts (i.e., those transcribed by A) not available in the current ArchiMob release.

The training set contains about 14,000 instances (between 3,000 and 4,000 instances per dialect) with a total of 114,000 tokens (28,000 per dialect). The test set contains about 3,600 instances (900 per dialect) with a total of 29,500 tokens (7,000–8,000 per dialect). We did not provide a development set. The acoustic data were not released in this edition, but they are in principle available.

4.2 Task Setup

The task setup of the German Dialect Identification (GDI) task was analogous to the DSL task, except that we did not allow open training, because the test sets for the Zurich and the Bern dialects were already made publicly available through the ArchiMob release.

4.3 Participants and Approaches

A total of ten teams participated in the GDI task, which is very close to the participation in this year's DSL task (11 teams), but somewhat lower than the first edition of ADI (18 teams). All teams except one (*CLUZH*) also participated in the DSL or the ADI tasks. Below, we provide a short description of the approach taken by each team, where the teams are ordered by their performance on the test data in descending order:

- **MAZA** This team submitted three runs, all of which are based on a combination of probabilistic classifiers. Their best run (run 3) is a meta-classifier based on individual SVM classifiers using character 1–8-grams and word unigrams (Malmasi and Zampieri, 2017b).
- **CECL** This team submitted three runs, all based on SVM classifiers using character 1–5-grams, weighted by BM25. The different runs used different decision rules, with run 3 performing best (Bestgen, 2017).
- **CLUZH** This team submitted three runs. Run 1 used a Multinomial Na"ive Bayes classifier with character n-grams. Run 2, which performed best, used a Conditional Random Fields (CRF) classifier, where each word of the sentence is represented by character n-gram features, prefix and suffix n-gram combinations, and word shapes. Run 3 used majority voting of runs 1 and 2, and an SVM classifier (Clematide and Makarov, 2017).
- **qcri_mit** This team submitted three runs based on different combinations of SVM classifiers and Stochastic Gradient classifiers with different loss functions. Their best-performing run (run 3) consisted of an SVM classifier with 1–5-grams, another SVM with 1–8-grams, and an SGD with Modified Huber Loss and L2 regularization and 1–5-gram features.

- **unibuckernel** This team submitted three runs, all of which are based on multiple string kernels combined with either Kernel Ridge Regression (KRR) or Kernel Discriminant Analysis. Their best run (run 1) used a KRR classifier trained on the sum of the blended presence bits kernel based on 3–6-grams, the blended intersection kernel based on 3–6-grams, and the kernel based on LRD with 3–5-grams (Ionescu and Butnaru, 2017).

- **tubasfs** This team submitted a single system, based on a linear SVM classifier. Their system used both characters and words as features, and optimized hyperparameters (the n-gram size and margin/regularization parameter for SVM) (Çöltekin and Rama, 2017).

- **ahaqst** This team submitted two runs, both based on cross-entropy. Run 2, which performed better, approximated cross-entropy using strings of up to 25 bytes (Hanani et al., 2017).

- **Citius_Ixa_Imaxin** This team submitted three runs, all of which are based on language model perplexity. Run 2 was based on word unigram features, and it was their best (Gamallo et al., 2017).

- **XAC_Bayesline** This team submitted one run. As for DSL, it is an adaptation of the system submitted to the DSL 2016 by Barbaresi (2016).

- **deepCybErNet** This team submitted two runs based on LSTM neural networks. Run 1 uses character features, whereas run 2 uses word features.

4.4 Results

Table 7 shows the results of the GDI task, reporting the best run of each team. Like in the DSL task, all teams except *deepCybErNet* obtained similar scores.

The per-dialect results look rather similar across the teams. For BE and BS, precision and recall were fairly balanced around 0.7. LU is characterized by very low recall (around 0.3), whereas ZH features higher than average recall values of around 0.9. An exception to this trend is the *CECL* submission, which shows more balanced figures for LU, with a recall of 0.52, but at the expense of precision: 0.55 instead of around 0.7.

Rank	Team	F1 (weighted)
1	MAZA	0.662
2	CECL	0.661
3	CLUZH	0.653
4	qcri_mit	0.639
5	unibuckernel	0.637
6	tubasfs	0.626
7	ahaqst	0.614
8	Citius_Ixa_Imaxin	0.612
9	XAC_Bayesline	0.605
10	deepCybErNet	0.263

Table 7: GDI task: closed submission results.

The bad performance of LU can be explained by transcriber effects. As shown in Table 6, it is the only dialect for which no utterances from the test transcriber (A) are included in the training set. This hypothesis is supported by the fact that LU is most often confused with BS (which contains training data by A, but is dialectologically rather distant from LU), and by the fact that the participants have not observed such low recall in their cross-validation experiments on the training data. The exact nature of these transcriber effects remains to be investigated and should be better controlled in future iterations of this shared task.

We see two reasons for the high recall of ZH. On the one hand, the training set is dialectally more homogeneous (all documents except for one stem from the city of Zurich and its suburbs) but more heterogeneous in terms of document and transcriber distributions. This probably allows the models to focus on dialectal specificities and to disregard spurious transcriber particularities. On the other hand, Scherrer and Rambow (2010) as well as Hollenstein and Aepli (2015) found ZH to be one of the most easily identifiable dialects, suggesting that it acts as a sort of default dialect with few characteristic traits. Dialectometrical studies (Scherrer and Stoeckle, 2016) have partially confirmed this role of the Zurich dialect.

4.5 Summary

This first edition of the GDI task was a success, given the short time between the 2016 and 2017 editions. In the future, we would like to better control transcriber effects, either by a more thorough selection of training and test data, or by adding transcriber-independent features such as acoustic features, as has been done in the ADI task this year. Further dialectal areas could also be added.

5 Cross-lingual Dependency Parsing (CLP)

VarDial 2017 featured for the first time a cross-lingual parsing task for closely related languages.[7] Transfer learning and annotation projection are popular approaches in this field and various techniques and models have been proposed in the literature in particular in connection with dependency parsing (Hwa et al., 2005; McDonald et al., 2013; Täckström et al., 2012; Tiedemann, 2014). The motivation for cross-lingual models is the attempt to bootstrap tools for languages that do not have annotated resources, which are typically necessary for supervised data-driven techniques, using data and resources from other languages. This is especially successful for closely related languages with similar syntactic structures and strong lexical overlap (Agić et al., 2012). With this background, it is a natural extension for our shared task to consider cross-lingual parsing as well. We do so by simulating the resource-poor situation by selecting language pairs from the Universal Dependencies (UD) project (Nivre et al., 2016) that match the setup and come close to a realistic case for the approach (using UD release 1.4). The UD datasets are especially useful as they try to harmonize the annotation across languages as much as possible, which facilitates the cross-lingual scenario.

Language	Sentences	Words
Czech	68,495	1.3M
Danish	4,868	89k
Swedish	4,303	67k
Slovenian	6,471	119k

Table 8: CLP task: source language training data.

We selected Croatian, Norwegian and Slovak as the target languages and pre-defined source languages that may be used for the cross-lingual parsing. For Norwegian, we have two possible source languages: Danish and Swedish. For Croatian, the source is Slovenian, and for Slovak it is Czech. We provided training data for each source language (a copy of the original UD data), pre-trained part-of-speech (PoS) and morphological taggers for the target languages, and development data with predicted PoS labels and predicted morphology (based on the provided taggers).

[7]For data and other information see `https://bitbucket.org/hy-crossNLP/vardial2017`

Avoiding gold labels is important here in order to avoid exaggerated results that blur the picture of a more realistic setup (Tiedemann, 2015). The tagger models are trained on the original target language treebanks using UDpipe (Straka et al., 2016) with standard settings and without any optimization of the hyper parameters. The size of the source language data is given in Table 5. We can see that for Czech we have by far the largest corpus, which will also be reflected in the results we obtain.

Language-pair	Sentences	Words
Czech-Slovak	5.7M	77M
Danish-Norwegian	4.9M	69M
Swedish-Norwegian	4.2M	60M
Slovenian-Croatian	12.8M	172M

Table 9: CLP task: parallel training data.

Participants were asked not to use the development data with their gold standard annotation of dependency relations for any training purposes. The purpose of the development datasets is entirely for testing model performance during system development. All the knowledge used for parsing should origin in the provided source language data. Other sources (except for target language sources) could also be used in unconstrained submissions, but none of the participants chose that option. For the constrained setup, we also provided parallel datasets coming from OPUS (Tiedemann, 2012) that could be used for training cross-lingual parsers in any way. The datasets included translated movie subtitles and contained quite a bit of noise in terms of alignment, encoding, and translation quality. They were also from a very different domain, which made the setup quite realistic considering that one would used whatever could be found for the task. The sizes of the parallel datasets are given in Table 8.

In the setup of the shared task, we also provided simple baselines and an "upper bound" of a model trained on annotated target language data. The cross-lingual baselines included delexicalized models (based on universal PoS tags only) and a straightforward application of lexicalized source language parsers to the target language without any kind of adaptation. All these models were trained using UDPipe without any parameter optimization and should be seen as lazy baselines for rapid tool development.

Supervised Models		LAS	UAS
Croatian	Croatian	68.51	75.61
Norwegian	Norwegian	78.23	82.28
Slovak	Slovak	69.14	76.57
Delexicalized Models		LAS	UAS
Croatian	Slovenian	50.81	62.64
Norwegian	Danish	55.17	65.23
Norwegian	Swedish	57.54	66.96
Norwegian	Danish+Swedish	58.80	68.58
Slovak	Czech	48.91	60.68
Non-adapted Source Models		LAS	UAS
Croatian	Slovenian	53.35	63.94
Norwegian	Danish	54.91	64.53
Norwegian	Swedish	56.63	66.24
Norwegian	Danish+Swedish	59.95	69.02
Slovak	Czech	53.72	65.70

Table 10: CLP task: baseline models in terms of labeled attachment scores (LAS) and unlabeled attachment scores (UAS).

We received three submissions (denoted by *tubasfs*, *CUNI* and *Helsinki-CLP*) for the CLP task and all of them submitted results for all language pairs. All three submissions used some kind of annotation projection instead of model transfer. Two of them applied word-by-word translation (Çöltekin and Rama, 2017; Rosa et al., 2017) based on lexical translations learned from the parallel corpora. The third one (Tiedemann, 2017) applied a mix of annotation projection (Tiedemann, 2014) and treebank translation (Tiedemann et al., 2014). The overall results are shown in Table 11.

LAS	**Croatian**	**Norwegian**	**Slovak**
CUNI	60.70	70.21	78.12
Helsinki-CLP	57.98	68.60	73.14
tubasfs	55.20	65.62	64.05
UAS	**Croatian**	**Norwegian**	**Slovak**
CUNI	69.73	77.13	84.92
Helsinki-CLP	69.57	76.77	82.87
tubasfs	75.61	74.61	73.16

Table 11: CLP task: closed submission results.

From the results, we can see that *CUNI* is the clear winner especially in terms of labeled attachment scores. The difference to the second-best submission is large in particular on the Slovak data. The picture is not that clear in terms of unlabeled attachment scores.

The difference in LAS between the two top submissions is most likely due to the label normalization that the winning system applied besides the direct annotation projection. They also applied a more selective projection of morphological features and used the extensive parallel data provided for the task in order to train reliable word embeddings for the target language. Another improvement was obtained by relabeling the test sets with morpho-syntactic information learned from the projected datasets. This is especially useful for Slovak, which gains a lot from the tagger that is trained on large amounts of projected Czech data instead of applying the information provided by the supervised tagger trained on smaller amounts of target language data. Their system also applied a joint model for tagging and parsing, which improved the overall performance.

We can also see striking differences between the results for the three target languages. Overall, Croatian is the least successful case with improvements of 2-7 points in LAS over the non-adapted baseline. For Norwegian, the two top-scoring teams achieve over 10 LAS points of improvement for the winning submission. However, for both Croatian and Norwegian, the cross-lingual models are still far behind the fully-supervised upper bound that scores 8 LAS points above them. For Slovak, the picture is different. The two top submissions both score above the "upper bound" of fully-supervised parsing, which is quite an impressive result. This is certainly due to the large amounts of training data that we have for the source language (Czech) and the close relation between the two languages supports the success as well. Nevertheless, the results demonstrate the real-world use of the techniques tested in our shared task.

6 Conclusion and Future Work

We have presented the methods, the data, the evaluation setup, and the results for four shared tasks taht we organized as part of the VarDial 2017 evaluation campaign. To the best of our knowledge, this is the first comprehensive evaluation campaign on NLP for Similar Languages, Varieties and Dialects. Three tasks (ADI, GDI, and DSL) dealt with dialect and language variety identification, focusing on Arabic, German and several groups of similar languages, respectively, whereas the CLP task dealt with parsing.

Along with the results of each shared task, we also included short descriptions of each participating system in order to provide readers with an overview of all approaches proposed for each task. For a complete description of each system, we included references to the fifteen system description papers that were accepted for presentation at the VarDial workshop at EACL'2017.

Given the success of the VarDial evaluation campaign, we believe that there is room for another edition with more shared tasks. Possible topics of interest for future shared tasks include machine translation between similar languages and POS tagging of dialects, among others.

Acknowledgments

We would like to thank the participants of the previous editions of the DSL shared task for their participation, support, and feedback, which have motivated us to turn the task into a more comprehensive evaluation campaign as of this year.

We further thank all participants in this year's VarDial evaluation campaign for their valuable comments and suggestions.

References

Kheireddine Abainia, Siham Ouamour, and Halim Sayoud. 2016. Effective Language Identification of Forum Texts Based on Statistical Approaches. *Information Processing & Management*, 52(4):491–512.

Željko Agić, Danijela Merkler, and Daša Berović. 2012. Slovene-Croatian treebank transfer using bilingual lexicon improves Croatian dependency parsing. In *Proceedings of IS-LTC*.

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and Jim Glass. 2014a. A complete Kaldi recipe for building Arabic speech recognition systems. In *Proceedings of SLT*.

Ahmed Ali, Yifan Zhang, and Stephan Vogel. 2014b. QCRI Advanced Transcription System (QATS). In *Proceedings of SLT*.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016. Automatic dialect detection in Arabic broadcast speech. In *Proceedings of INTERSPEECH*.

Mohamad Hasan Bahari, Najim Dehak, Lukas Burget, Ahmed Ali, Jim Glass, et al. 2014. Non-negative factor analysis for GMM weight adaptation. *IEEE Transactions on Audio Speech and Language Processing*.

Adrien Barbaresi. 2016. An unsupervised morphological criterion for discriminating similar languages. In *Proceedings of the VarDial Workshop*.

Adrien Barbaresi. 2017. Discriminating between similar languages using weighted subword features. In *Proceedings of the VarDial Workshop*.

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the VarDial Workshop*.

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of EMNLP*.

Dayvid Castro, Ellen Souza, and Adriano LI de Oliveira. 2016. Discriminating between Brazilian and European Portuguese national varieties on Twitter texts. In *Proceedings of BRACIS*.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear SVMs and neural networks. In *Proceedings of the VarDial Workshop*.

Çağrı Çöltekin and Taraka Rama. 2017. Tübingen system in VarDial 2017 shared task: Experiments with language identification and cross-lingual parsing. In *Proceedings of the VarDial Workshop*.

Simon Clematide and Peter Makarov. 2017. CLUZH at VarDial GDI 2017: Testing a variety of machine learning tools for the classification of Swiss German dialects. In *Proceedings of the VarDial Workshop*.

Marcelo Criscuolo and Sandra Aluisio. 2017. Discriminating between similar languages with word-level convolutional neural networks. In *Proceedings of the VarDial Workshop*.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift*. Sauerländer, Aarau, 2 edition.

Pablo Gamallo, Jose Ramon Pichel, and Iñaki Alegria. 2017. A method based on perplexity for similar languages discrimination. In *Proceedings of the VarDial Workshop*.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with TF-IDF weighting. In *Proceedings of the BEA workshop*.

Helena Gómez-Adorno, Ilia Markov, Jorge Baptista, Grigori Sidorov, and David Pinto. 2017. Discriminating between similar languages using a combination of typed and untyped character n-grams and words. In *Proceedings of the VarDial Workshop*.

Cyril Goutte and Serge Léger. 2015. Experiments in discriminating similar languages. In *Proceedings of the LT4VarDial Workshop*.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2013. Feature space selection and combination for native language identification. In *Proceedings of the BEA Workshop*.

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

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of LREC*.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the W-NUT Workshop*.

Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor. 2017. Identifying dialects with textual and acoustic cues. In *Proceedings of the VarDial Workshop*.

Nora Hollenstein and Noëmi Aepli. 2015. A resource for natural language processing of Swiss German dialects. In *Proceedings of GSCL*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Radu Tudor Ionescu and Andrei Butnaru. 2017. Learning to identify Arabic and German dialects using multiple kernels. In *Proceedings of the VarDial Workshop*.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the BEA Workshop*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2015. Discriminating Similar Languages with Token-based Backoff. In *Proceedings of the LT4VarDial Workshop*.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. HeLI, a word-based backoff method for language identification. In *Proceedings of the VarDial Workshop*.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017. Evaluating HeLI with non-linear mappings. In *Proceedings of the VarDial Workshop*.

Sameer Khurana and Ahmed Ali. 2016. QCRI advanced transcription system (QATS) for the Arabic Multi-Dialect Broadcast Media Recognition: MGB-2 Challenge. In *Proceedings of SLT*.

Adrian Leemann, Marie-José Kolly, Ross Purves, David Britain, and Elvira Glaser. 2016. Crowdsourcing language change with smartphone applications. *PLOS ONE*, 11(1):1–25.

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

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of ALTA*.

Marco Lui. 2014. *Generalized language identification*. Ph.D. thesis, University of Melbourne.

Shervin Malmasi and Mark Dras. 2015a. Automatic language identification for Persian and Dari texts. In *Proceedings of PACLING*.

Shervin Malmasi and Mark Dras. 2015b. Language identification using classifier ensembles. In *Proceedings of the VarDial Workshop*.

Shervin Malmasi and Marcos Zampieri. 2017a. Arabic dialect identification using iVectors and ASR transcripts. In *Proceedings of the VarDial Workshop*.

Shervin Malmasi and Marcos Zampieri. 2017b. German dialect identification in interview transcriptions. In *Proceedings of the VarDial Workshop*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the VarDial Workshop*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*.

Maria Medvedeva, Martin Kroon, and Barbara Plank. 2017. When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In *Proceedings of the VarDial Workshop*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Will Radford and Matthias Gallé. 2016. Discriminating between similar languages in Twitter using label propagation. *arXiv preprint arXiv:1607.05408*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. In *Proceedings of TREC*.

Rudolf Rosa, Daniel Zeman, David Mareček, and Zdeněk Žabokrtský. 2017. Slavic forest, Norwegian wood. In *Proceedings of the VarDial Workshop*.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob – a corpus of spoken Swiss German. In *Proceedings of LREC*.

Upendra Sapkota, Steven Bethard, Manuel Montes-y Gómez, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of NAACL*.

Carolina Scarton, Liling Tan, and Lucia Specia. 2015. USHEF and USAAR-USHEF participation in the WMT15 quality estimation shared task. In *Proceedings of WMT*.

Yves Scherrer and Owen Rambow. 2010. Word-based dialect identification with georeferenced rules. In *Proceedings of EMNLP*.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP*.

Milan Straka, Jan Hajič, and Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of LREC*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of NAACL*.

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 BUCC Workshop*.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of CoNLL*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of LREC*.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING*.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *Proceedings of Depling*.

Jörg Tiedemann. 2017. Cross-lingual dependency parsing for closely related languages - Helsinki's submission to VarDial 2017. In *Proceedings of the VarDial Workshop*.

Lan Wang, Masahiro Tanaka, and Hayato Yamana. 2016. What is your Mother Tongue?: Improving Chinese native language identification by cleaning noisy data and adopting BM25. In *Proceedings of ICBDA*.

Samantha Wray and Ahmed Ali. 2015. Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic. In *Proceedings of INTERSPEECH*.

Fan Xu, Mingwen Wang, and Maoxi Li. 2016. Sentence-level dialects identification in the Greater China region. *International Journal on Natural Language Computing (IJNLC)*, 5(6).

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. N-gram language models and POS distribution for the identification of Spanish varieties.

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

Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. 2015a. Comparing approaches to the identification of similar languages. In *Proceedings of the VarDial Workshop*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015b. Overview of the DSL shared task 2015. In *Proceedings of the LT4VarDial Workshop*.

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 TweetLID Workshop*.