# Discriminating between Similar Languages on Imbalanced Conversational Texts

**Junqing He[†‡], Xian Huang[§], Xuemin Zhao[†‡], Yan Zhang[†‡], Yonghong Yan[†‡♣]**
[†]Institute of Acoustics, Chinese Academy of Science, Bejing, China
[‡] University of Chinese Acadamy of Sciences, Bejing, China
[§]PLA Information Engineering University Luoyang Division, Henan, China
[♣]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumchi, China
hejunqing@hccl.ioa.ac.cn, a77huang852yi@126.com, {zhaoxuemin,zhangyan,yanyonghong}@hccl.ioa.ac.cn

**Abstract**

Discriminating between similar languages (DSL) on conversational texts is a challenging task. This paper aims at discriminating between limited-resource languages on short conversational texts, like Uyghur and Kazakh. Considering that Uyghur and Kazakh data are severely imbalanced, we leverage an effective compensation strategy to build a balanced Uyghur and Kazakh corpus. Then we construct a maximum entropy classifier based on morphological features to discriminate between the two languages and investigate the contribution of each feature. Empirical results suggest that our system achieves an accuracy of 95.7% on our Uyghur and Kazakh dataset, which is higher than that of the CNN classifier. We also apply our system to the out-of-domain subtasks of VarDial'2016 DSL shared tasks to test the system's performance on short conversational texts of other similar languages. Though with much less preprocessing, our system outperforms the champions on both test sets B1 and B2.

**Keywords:** discriminating between similar languages, imbalanced data, morphological features, maximum entropy

## 1. Introduction

Automatic language identification (LID) aims to identify the language a document is written in, which is an important branch in Natural Language Processing (NLP) (Zampieri et al., 2015a). The past two decades had witnessed fast development in LID and state-of-the-art systems have achieved high accuracy (Simões et al., 2014) and wide coverage (Brown, 2014) on standard texts. However, identifying languages from very little data, from multi-languages input or discriminating between extremely similar languages are bottlenecks of this field (Ljubešic and Kranjcic, 2014; Zampieri et al., 2015b). What's more, identifying similar languages with limited resource is unsolved.

Uyghur and Kazakh, widely used in Middle East and North West of China, are similar languages. They both belong to the Turkic group of Altaic family and are agglutinative languages. According to Wang et al. (2013), the similarity between Uyghur and Kazakh at sentence and word level are over 80% and 90% respectively. They have many characteristics in common: (1) They are both written in Arabic alphabets in the right-to-left order. (2) Theoretically there are 32 letters in Uyghur and 33 in Kazakh. The two languages share 26 letters and encoding areas with another 2 letters look exactly the same. (3) There is a large overlap of vocabulary and syntax between the two languages. It is very difficult to identify them by looking up the words in dictionaries. (4) In both languages, a great amount of prefixes and suffixes are attached to a word, which makes word stemming and recognition difficult.

Here we define "short conversational texts" as short texts people used to communicate with each other through mobile devices, communicational software and social-media platforms. They can be (1) short messages people send to each other through cell phones; (2) chatting records of communicational software such as Wechat and MSN; (3) posts and comments on social-media platforms such as Twitter, Facebook and Microblog. These texts are obstacles for NLP tasks for the following reasons: (1) Each text is pretty short. Lengths of most sentences range from 3 to 9 words. (2) There exist enormous spelling and grammatical mistakes in the texts, which make it time and energy consuming in word stemming and error correction. (3) Abbreviations and colloquial expressions are widely used. (4) It takes much time and energy to collect short conversational texts, resulting in the imbalance and inadequacy of the corpus. (5) Without unified input methods, people use various characters other than standard ones. In fact, more than 100 letters of different encoded bytes are found in our corpus. This strengthened the difficulty of discriminating between Uyghur and Kazakh short conversational texts.

## 2. Related Work

Since more and more researchers are concerned with discriminating between similar languages (DSL), a series of shared tasks were organized by the workshop series for Similar Languages, Varieties and Dialects (VarDial), which was collocated with either COLING, RANLP or EACL. According to Malmasi et al. (2016), high-order character n-grams were the most successful feature, and the best classification models included SVM, logistic regression, and language models, while deep learning approaches did not perform very well.

To deal with short and sparse texts, solutions (Phan et al., 2008; Rehurek and Kolkus, 2009; Tromp and Pechenizkiy, 2011; Dai et al., 2013) were proposed to enrich short text representation by bringing in additional semantics. The additional semantics could be from data collection itself or be derived from a much larger external knowledge base. Dealing with tweets, Zubiaga et al. (2014) summarized the TweetLID shared task and workshop held at SEPLN 2014 and pointed out several shortcomings in current researches.

When it comes to discriminating between Uyghur and Kazakh, Hasimu et al. (2015) employed unique characters to identify the Uyghur , Kazakh and Kyrgyz languages. They carried out experiments on the written texts longer than 70 words and achieved a 96.67% accuracy. But in the web corpus of less than 10 words, the precision of Kazakh fell dramatically to only 65.31%.

In this paper, we made two contributions: (1) We constructed a corpus of conversational texts in Uyghur and Kazakh for similar languages identification and proposed a method for corpus augmentation. (2) We designed a system that can effectively discriminate between similar languages on conversational texts.

## 3. Data Construction

### 3.1. Data Collection

With the popularization of social network and chatting applications on mobile phones, people are more likely to communicate with each other via short instant messages. Thus natural languge processing on short conversational messages is of great significance.

We collected 48680 texts from the chatting messages sent by mobile phones and used as our training set after anonymization. Likewise, 973 colloquial messages that were sent in a day were collected as our test set. Then all the texts were tagged by linguistic experts. In our training set, we found that 48432 samples were written in Uyghur while 148 samples were Kazakh. As for test set, 687 and 286 texts were annotated as Uyghur and Kazakh separately. The scales of Uyghur and Kazakh texts in the training set were severely imbalanced, which exceeded the proportion of 327:1.

### 3.2. Data Augmentation

Since a highly imbalanced training corpus may hinder the effectiveness of discrimination between the two similar languages, we decided to balance the corpus by supplementing Kazakh texts.

We did not collect more Kazakh short messages in the same way because of inefficiency since the linguistic experts have to skim more than 300 Uyghur samples to get a Kazakh sample. To obtain data that are similar to short messages which are conversational, informal and short, we decided to crawl data from Kazakh forums instead of the Kazakh news web pages and Twitter. The reasons are as follows: (1) News are formal written texts, which have little overlap of words and characters with short communicational messages. (2) Although tweets are short and informal, Twitter is rarely used in China. (3) Posts on Kazakh forums are informal and conversational, which resemble the nature of short messages. What's more, the contents are almost entirely written in Kazakh.

We crawled 70909 web pages from a Kazakh forum[1]. However, some texts in these web pages were longer than the chatting messages. To make the crawled texts more similar to the short messages, we picked out 339,609 samples of no more than 14 words. Then we randomly chose 48000 texts from the filtered samples to match the number of Uyghur texts in the original training set. In this way, we constructed a balanced corpus of Uyghur and Kazakh short conversational texts used for discriminating between the two similar languages. The final corpus contains a training set with 48432 Uyghur and 48148 Kazakh samples; and and a test set with 687 Uyghur and 286 Kazakh samples. More details of the corpus are listed in Table 1 and Table 2.

[1]From `http://bbs.senkazakh.com`

| | Training Set | | Test Set | |
|---|---|---|---|---|
| Language | Uyghur | Kazakh | Uyghur | Kazakh |
| Vocab. Size | 37237 | 43876 | 2528 | 1161 |
| Instances | 48432 | 48148 | 687 | 286 |
| Avg. Length | 5.98 | 4.39 | 5.94 | 5.77 |

Table 1: Statistics of the training and test set of the Uyghur and Kazakh data. Avg. Length represents the average number of tokens in each instance.

| Num. of Tokens | Test Set | Training Set | | |
|---|---|---|---|---|
| | | Uyghur | Kazakh | Sup. Kazakh |
| 1 | 25 | 135 | 21 | 4032 |
| 2 | 93 | 1433 | 9 | 7907 |
| 3 | 135 | 4847 | 16 | 8558 |
| 4 | 116 | 7668 | 13 | 7755 |
| 5 | 88 | 8170 | 17 | 5688 |
| 6 | 74 | 7352 | 10 | 4359 |
| 7 | 81 | 6227 | 12 | 3637 |
| 8 | 158 | 5409 | 17 | 2751 |
| 9 | 118 | 3898 | 18 | 1856 |
| 10 | 63 | 2299 | 10 | 1046 |
| 11 | 19 | 837 | 4 | 341 |
| 12 | 3 | 198 | 1 | 51 |
| 13 | 0 | 42 | 0 | 9 |
| 14 | 0 | 7 | 0 | 1 |
| Sum | 973 | 48432 | 148 | 48000 |

Table 2: Distributions of instances on different number of tokens in the Uyghur and Kazakh data for DSL. Sup.Kazakh is referred to the supplemental Kazakh data .

## 4. Our System

### 4.1. Feature Extraction

Since all texts in the corpus are extremely short, we assume the lexical n-gram features cannot play an important role in DSL in short conversational texts. Based on linguistic, in particular morphological analysis of the two languages, we mainly used the following features to discriminate between the two languages:

- **Unique characters**. Once a unique character is found in a text, we can determine that the text is written in the language the unique character belongs to.
- **Character n-grams**. The sequence and combination of characters is different among various languages, even though the languages share a lot of characters. The Uyghur Latin word "men" corresponds to "man" in Kazakh Latin for the same meaning. (For the convenience of typing and visualization, here we use Latin

charcters to embody the Uyghur and Kazakh instead of the Arabic letters.)

- **Prefixes and suffixes**. As agglutinative languages, both Uyghur and Kazakh have numerous affixes. On many occasions, affixes of the two languages are different. For example, to express the same meaning, suffix "lar" is used in Uyghur, while "dar" is used in Kazakh. Likewise, "o" can be the first letter in Kazakh but cannot be found at this position in Uyghur. One thing we should note is that misspelling problems make this feature hard to extract, thus we use the the first and last n characters of the words as a substitute of prefixes and suffixes. Here n ranges from 1 to 3.

- **Word unigrams**. The frequency of a word represents how likely it belongs to a language. If a text contains a high-frequency word of a language, it is more likely to belong to the corresponding language.

- **Bin on text length**. We can divide the texts into different bins according to the lengths of the texts. Models trained in certain bin length will be more accurate.

### 4.2. Classifiers

Nowadays there are many state-of-the-art classifiers that achieve steady and desirable performance, no matter whether they are based on machine learning or neural networks.

The maximum entropy (MaxEnt) classifier is one of the best models among the machine learning algorithms. The MaxEnt classifier computes the conditional likelihood and relativity of the features mutually for each category in the training step. Based on the statistics, for each sample, the classifier adjusts the weights of corresponding features to maximize the max entropy of the sentence under the constraints of all the conditional likelihood above. When predicting, scores of samples of each category is computed and the class of the highest score is chosen as its label. Therefore, feature dependence is taken into account in MaxEnt. In this paper, we applied a MaxEnt classifier in our system using the Stanford classifier toolkit[2].

With the convolutional neural networks (CNN) successfully applied to image recognition (Krizhevsky et al., 2012) and text classification (Kim, 2014), CNN became one of the most popular deep learning classifier algorithms. We also built a CNN classifier based on character embeddings considering the features are mainly of the character level and then compared the performances of the two classifiers.

### 4.3. Evaluation

Since we take the DSL task as an issue of classification, we use the evaluation metrics of classification systems. Precision (P), recall (R) and accuracy (Acc) are used to evaluate the performance of our system.

## 5. Experiments and Discussion

In this section, we conducted four experiments to examine the effect of the supplemented Kazakh samples, the contribution of each morphological feature we use, the performance of the CNN and MaxEnt classifiers, and the performance of our system in dealing with the out-of-domain test sets B1, B2 of the VarDial'2016 DSL shared tasks. The B1 and B2 data sets are considered to be out-of-domain because the training data are collected from news while the test sets comprise of tweets.

### 5.1. Experiment on the Supplemented Data

In this experiment, we use the MaxEnt classifier based on all the features except bin. Table 3 shows the results of our system trained on the original imbalanced and the final supplemented training sets.

| Training Sets | Uyghur | | Kazakh | | |
|---|---|---|---|---|---|
| | P | R | P | R | Acc |
| Original | 89.0 | 99.3 | 97.6 | 70.6 | 90.0 |
| Final | **98.5** | 95.1 | 89.0 | **96.5** | 95.5 |

Table 3: The influence of data augmentation. All the results are in percentage.

Results of the experiment reveal that the model trained on the imbalanced training data can recall almost all the Uyghur texts but can only recall 70.6% of true Kazakh samples. It is indicated that the system regards most of the samples as Uyghur with only 89.0 % of precision since the high probability of its appearance in the original training set. After the data augmentation, the recall of the Kazakh samples improves by 25.9% and is close to that of the Uyghur samples. It is suggested that the augmentation strategy is useful and it is important to keep the scales of the training data for each language even.

### 5.2. Experiment on the Features' Contribution

To investigate the contribution of each feature, we evaluate the performance of our system using all the features, and without each one of them each time separately, *e.g.* using all the features without unique characters or character n-grams. In this way, we can see how importance each feature is by observing the decrease of performance, compared to that of using all the features. The MaxEnt classifier is employed and trained on the final Uyghur and Kazakh training set. Results are shown in Table 4.

| Features | Uyghur | | Kazakh | | |
|---|---|---|---|---|---|
| | P | R | P | R | Acc |
| All | 98.5 | 95.2 | 89.3 | 96.5 | 95.6 |
| -unique chars | 98.3 | 94.5 | 87.9 | 96.2 | 95.0 |
| -char n-gram | 97.8 | 90.5 | 80.7 | 95.1 | 91.9 |
| -pre/suf-fixs | 98.5 | 94.6 | 88.2 | 96.5 | 95.2 |
| -word unigrams | **98.5** | **95.3** | **89.6** | **96.5** | **95.7** |
| -bin | 98.5 | 95.0 | 89.0 | 96.5 | 95.5 |

Table 4: Performance of our system without each kind of features. All the results are in percentage. The bold results are the best performance in the same metric.

As we can see, with each feature removed, the performance of our system decreases to different extent except

[2]Available at `https://nlp.stanford.edu/software/`

for the word unigrams. It proves that all the features except word unigrams are useful in this task. We also observe the sharpest decline in accuracy when the character n-grams are removed, which implies it contributes the most among all the features. On the contrast, the accuracy increases slightly without the word unigrams, which reveals that word level features are helpless and even undermine the performance of language identification on the data set. Therefore, we stop using this feature in the following experiments.

## 5.3. Experiment on the Classifiers

In this experiment we respectively use the CNN classifier and MaxEnt classifier trained on the final Uyghur and Kazakh corpus to compare their performance. For the MaxEnt classifier, all the features except word unigrams are used. For the CNN classifier, the samples are represented at the character level with each character mapped into an embedding of 50 dimensions. Convolutional kernel widths are set to [1,2,3,4] to resemble the character n-grams of size 1 to 4 used in MaxEnt classifier. Numbers of kernels are set to be [50,200,300,500] separately since there is no improvement when using more kernels. A dropout layer with a 0.5 dropout rate is applied. The character embeddings are randomly initialized between (-0.05,0.05) under the uniform distribution. The performances of the two classifiers using the best parameters are listed in Table 5.

| Classifiers | Uyghur | | Kazakh | | |
|---|---|---|---|---|---|
| | P | R | P | R | Acc |
| CNN | 30.7 | 10.1 | 71.4 | 93.6 | 69.1 |
| MaxEnt | 98.5 | 95.3 | 89.6 | 96.5 | 95.7 |

Table 5: Performances of the MaxEnt and CNN classifier. All results are in percentage.

As Table 5 indicates, the MaxEnt classifier turned out to be much more competitive and effective than the CNN classifier. This finding echoes Zampieri et al. (2017) and Malmasi et al. (2016)'s findings that CNN fails to perform well in DSL tasks. Having an insight into the CNN's performance on Uyghur and Kazakh samples, we can see much better results on Kazakh compared to that on Uyghur samples. According to Table 1, we assume that the CNN classifier needs more training data to learn the character n-gram patterns in Uyghur than in Kazakh samples since they have longer sentences. This lack of Uyghur training data may lead to the failure in identifying Uyghur samples.

## 5.4. Experiment on the VarDial'2016 DSL shared task

Since the MaxEnt classifier using the morphological features achieved a high accuracy in discriminating between Ughur and Kazakh, we intended to test the performance of our system in discriminating other similar languages on short conversational texts.

### 5.4.1. Data Description

We chose the two social media data (B1 and B2) of Subtask 1 in VarDial'2016 DSL shared task[3] as the test materials. The training set consists of 18000 instances of journalistic data per language for training and 2000 instances for development. Each of the test sets includes 100 Twitter users' tweets per language. A varying number of tweets from a user are concatenated as a test sample (98.88 and 50.47 tweets per user for B1 and B2 in average separately). The two test sets cover two groups of closely-related languages : South-Slavic (Bosnian, Croatian, Serbian) and Portuguese (Brazilian and European). For each sample in the test set which contains five language/variants in a messed order, we have to find out which language it belongs to.

The reasons why we chose the two test sets are as following: (1) The test sets consists of Tweets, which are short, informal and conversational. (2) The test sets are out-of-domain data, which can test the classifier's robustness in handling out-of-domain data. (3) We can test whether the features are effective in discriminating between other similar languages than the Uyghur and Kazakh.

In the preprocessing process, no measures was taken to deal with the training set. As to the test sets we just removed the links, at-mentions and hash tags in them. Details of the training set and processed test sets B1 and B2 of subtask 1 are shown in Table 6.

| Data Sets | Language Varieties | Instances | Vocab. Size | Avg. Length |
|---|---|---|---|---|
| Train | Bosnian | 18000 | 77851 | 36.51 |
| | Croatian | 18000 | 82670 | 42.70 |
| | Serbian | 18000 | 74726 | 39.64 |
| | BP | 18000 | 44415 | 48.44 |
| | EP | 18000 | 39056 | 44.43 |
| Test B1 | Bosnian | 100 | 30418 | 1270.61 |
| | Croatian | 100 | 24754 | 966.33 |
| | Serbian | 100 | 31278 | 1219.20 |
| | BP | 100 | 16457 | 960.94 |
| | EP | 100 | 14878 | 843.29 |
| Test B2 | Bosnian | 100 | 25225 | 997.31 |
| | Croatian | 100 | 17811 | 613.59 |
| | Serbian | 100 | 23103 | 791.23 |
| | BP | 100 | 3922 | 121.22 |
| | EP | 100 | 2803 | 78.40 |

Table 6: Statistical analysis of Subtask 1 in VarDial'2016 DSL shared tasks. BP is short for Brazilian Portuguese and EP is short for European Portuguese.

### 5.4.2. Evaluation

In the DSL shared task, average accuracy (Acc) and macro-averaged F1-score (F1) were used as the official scores. Therefore we use the same metrics in this experiment. Since the DSL datesets of the subtask are balanced with the same number of examples for each language variety, we

[3]The dataset is version 3.0 of DSLCC, which is available at `http://ttg.uni-saarland.de/resources/DSLCC`

mainly use the average accuracy for comparison in the following subsection.

#### 5.4.3. Results and Discussion

We applied the MaxEnt classifier and with the character n-gram feature (n ranges from 1 to 7) to compare with other participant systems. Results of our system as well as the top participant systems in B1 and B2 in the VarDial'2016 DSL shared tasks are listed in Table 7 and 8 respectively.

| Team | Acc | F1 | Approach |
|---|---|---|---|
| Our system | **0.930** | **0.930** | MaxEnt with char n-grams (n=1-7) |
| GW_LT3 | 0.920 | 0.919 | Logistic Reg. with char/word n-grams |
| nrc | 0.914 | 0.913 | Two-stage SVM with char 6-grams |
| UniBucNLP | 0.898 | 0.897 | Logistic Reg. With word 1,2-grams |
| UPV_UA | 0.888 | 0.886 | String kernels and discriminant analysis |
| tubasfs | 0.862 | 0.860 | SVM with char n-grams (n=1-7) |

Table 7: Results of top systems and our system on B1 in subtask 1 of Vardial'2016 DSL task. The bold results are the best performance in the same metric.

| Team | Acc | F1 | Approach |
|---|---|---|---|
| Our system | **0.890** | **0.890** | MaxEnt with char n-grams (n=1-7) |
| GW_LT3 | 0.878 | 0.877 | Logistic Reg. with char/word n-grams |
| nrc | 0.878 | 0.913 | Two-stage SVM with char 6-grams |
| UPV_UA | 0.858 | 0.857 | String kernels and discriminant analysis |
| UniBucNLP | 0.838 | 0.897 | Logistic Reg. With word 1,2-grams |
| tubasfs | 0.822 | 0.818 | SVM with char n-grams (n=1-7) |

Table 8: Results of top systems and our system on B2 in subtask 1 of Vardial'2016 DSL task. The bold results are the best performance in the same metric.

As is shown in Table 7 and 8, GW_LT3 ranked first in the subtask of discriminating between similar languages on the tweets dataset. It used character n-gram (n=2-6) and word n-gram (n=1-3) with term-frequency weighting, and took many preprocessing measures. Our system outperforms it by 1.0% in B1 and 1.2% in B2 in accuracy. It is implied that, besides Uyghur and Kazakh, our system is also highly efficient in DSL tasks in other similar languages on short conversational texts. Compared with tubasfs, which also used character n-grams as a feature (n=1-7), the accuracies of our system in B1 and B2 are both 6.8% higher. This indicates that MaxEnt is better than SVM in this task. In addition, while our system achieved the accuracy of 95.7% on the Uyghur and Kazakh dataset, we just set n to be 1 to 4 in the character n-gram feature. When dealing with B1 and B2 test sets, we set n to be 1 to 7, and the accuracies we got were 93.0% and 89.0% respectively, which are lower than that we got in dealing with Uyghur and Kazakh. The reason for the unsatisfying result is that the training set of DSL 2016 subtask1 are journalistic news, which are different from short conversational texts to some extent. That can also show that when discriminating between similar languages on short conversational texts, contents in related forums is a better resource than news as the training data.

## 6. Conclusion

In this paper, we have constructed a corpus of short conversational Uyghur and Kazakh texts used for DSL. To solve the severe imbalance problem of the two languages with limited resource, we proposed a data augmentation method. That was to crawl a Kazakh forum and choose the materials which were short, informal as supplemental data. It is suggested that our augmentation strategy is effective and texts from forums are more suitable than news texts for the DSL task.

Then we designed a MaxEnt classifier with morphological features to discriminate between Uyghur and Kazakh conversational texts. Our empirical study shows that the character level features we exploited are helpful while employing the word unigrams led to worse performance. Experimental results also indicate that our system can not only discriminate between Uyghur and Kazakh on short conversational texts at a high accuracy of 95.7%, but also outperforms the state-of-the-art systems in DSL on Tweets with out-of-domain training data in the VarDial'2016 DSL task. It is also implied that CNN is not a competitive model for this task and the MaxEnt performs better than the SVM classifier using the same features.

## Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 61271426, U1536117, 11504406, 11590770-4, 11590771), the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA06030100, XDA06030500, XDA06040603), National 863 Program (No. 2015AA016306), National 973 Program (No. 2013CB329302) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 201230118-3).

## 7. Bibliographical References

Brown, R. D. (2014). Non-linear mapping for improved identification of 1300+ languages. In *EMNLP*, pages 627–632.

Dai, Z., Sun, A., and Liu, X.-Y. (2013). Crest: Cluster-based representation enrichment for short text classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 256–267. Springer.

Hasimu, M., Silamu, W., Mushajiang, W., and Youliwasi, N. (2015). Unique character based statistical language identification for uyghur,kazak and kyrgyz. *Journal of Chinese Information Processing*, 29(2):111–117.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105.

Ljubešic, N. and Kranjcic, D. (2014). Discriminating between very similar languages among twitter users. In *Proceedings of the Ninth Language Technologies Conference*, pages 90–94.

Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., and Tiedemann, J. (2016). Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the third VarDial Workshop*, Osaka, Japan.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.

Rehurek, R. and Kolkus, M. (2009). Language identification on the web: Extending the dictionary method. In *CICLing*, pages 357–368. Springer.

Simões, A., Almeida, J. J., and Byers, S. D. (2014). Language identification: a neural network approach. In *OASIcs-OpenAccess Series in Informatics*, volume 38. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Tromp, E. and Pechenizkiy, M. (2011). Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34.

Wang, L., Yidemucao, D., and Silamu, W. S. (2013). An investigation research on the similarity of uyghur kazakh kyrgyz and mongolian languages. *Journal of Chinese Information Processing*, 27(6):180–186.

Zampieri, M., Gebre, B. G., Costa, H., and van Genabith, J. (2015a). Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 66–72. Association for Computational Linguistics.

Zampieri, M., Tan, L., Ljubešic, N., Tiedemann, J., and Nakov, P. (2015b). Overview of the dsl shared task 2015. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*.

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the vardial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain, April. Association for Computational Linguistics.

Zubiaga, A., Vicente, I. S., Gamallo, P., Campos, J. R. P., Loinaz, I. A., Aranberri, N., Ezeiza, A., and Fresno-Fernández, V. (2014). Overview of tweetlid: Tweet language identification at SEPLN 2014. In *Proceedings of the Tweet Language Identification Workshop co-located with 30th Conference of the Spanish Society for Natural Language Processing, TweetLID@SEPLN 2014, Girona, Spain, September 16th, 2014.*, pages 1–11.