Exploring lexical and syntactic features for language variety identification

van der Lee, Chris; van den Bosch, Antal

published in
Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)
2017

document version
Publisher's PDF, also known as Version of record

Exploring Lexical and Syntactic Features for Language Variety Identification

Chris van der Lee
Dept. of Communication and Information Sciences
Tilburg University
Tilburg, The Netherlands
c.vdrlee@tilburguniversity.edu

Antal van den Bosch
Centre for Language Studies
Radboud University, Nijmegen, and
Meertens Institute, Amsterdam,
The Netherlands
antal.van.den.bosch@meertens.knaw.nl

Abstract

We present a method to discriminate between texts written in either the Netherlandic or the Flemish variant of the Dutch language. The method draws on a feature bundle representing text statistics, syntactic features, and word n-grams. Text statistics include average word length and sentence length, while syntactic features include ratios of function words and part-of-speech n-grams. The effectiveness of the classifier was measured by classifying Dutch subtitles developed for either Dutch or Flemish television. Several machine learning algorithms were compared as well as feature combination methods in order to find the optimal generalization performance. A machine-learning meta classifier based on AdaBoost attained the best F-score of 0.92.

1 Introduction

Language identification, the task of automatically determining the natural language used in a document, is considered to be an important first step for many applications. Automatically determining a document's language can be a fairly easy step in certain situations (McNamee, 2005). However, some bottlenecks have been identified which leaves language identification unsolved as yet. It has been argued and demonstrated that one of the main bottlenecks is distinguishing between similar languages (Tiedemann and Ljubešić, 2012). Languages that are closely related such as Croatian and Serbian or Indonesian and Malay are very similar in their spoken and their written forms, which makes it difficult for automated systems to accurately discriminate between them. Recently, some advances have been achieved in the automated distinction between closely related languages, largely due to the *Discriminating between Similar Languages* (DSL) shared task. In the DSL competitions accuracies of over 95% have been reported, mostly using character and word n-grams with various classification algorithms.

Despite the fact that the accuracy of systems discriminating between similar languages is increasing, there are still challenges when it comes to discriminating between varieties of the same language, e.g. Spanish from South America or Spain. It has been claimed that language variety identification is even more difficult than similar language identification (Goutte et al., 2016). Results in the DSL competitions support this claim: only one system was able to score slightly above the 50% baseline when distinguishing between British and American English (Zampieri et al., 2014).

This work is related to recent studies that applied text classification methods to discriminate between written texts in different language varieties or dialects (Lui and Cook, 2013; Maier and Gómez-Rodriguez, 2014; Malmasi and Dras, 2015; Malmasi et al., 2015; Zampieri et al., 2016). The aim of the current work is to explore lesser studied techniques and features that could be beneficial to the accuracy of language variety classifiers. As a case study, classifiers were built to discriminate between Netherlandic Dutch and Flemish Dutch subtitles.

2 Related work

2.1 Language varieties

Research on varieties of the same language is scarce and the existing body of research on the topic shows that discriminating between language varieties is an even bigger challenge compared to similar languages. Six systems were submitted

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 190–199,
Valencia, Spain, April 3, 2017.

in the 2014 DSL shared task to discriminate between British English and American English, and only one of those systems scored above the 50% baseline (Zampieri et al., 2014). However, it is possible that the poor results attained in the 2014 DSL shared task were due to problems in the data set. Some classifiers have been built outside of the DSL shared task with higher accuracy scores. Lui and Cook (2013) built a classifier to distinguish the British, Canadian, and Australian English language varieties and tested this classifier on various corpora. The obtained F-scores varied greatly between the corpora: an F-score of over .9 was obtained in the best case, but scores were below the baseline in the worst cases.

Not only English language varieties have been studied. Maier and Gómez-Rodriguez (2014) developed a classifier to discriminate between five Spanish languages with tweets (short messages posted on the Twitter.com social media platform) as input. They achieved an average F-score of 0.34, which is somewhat above baseline, though not particularly high. Furthermore, Malmasi and Dras (2015) distinguished Dari and Farsi news texts with an accuracy of 96%. Malmasi et al. (2015) developed a classifier for multiple Arabic dialects. They achieved accuracy scores as high as 94%, but the results were relatively worse when they classified more closely-related dialects such as Palestinian and Jordanian (76%). Similarly, Zampieri et al. (2016) ventured to classify Portuguese news articles published in Brazil, Macau, and Portugal with differing accuracy scores. Macau versus European Portuguese was somewhat difficult (74%), while classifying Brazilian versus Macau Portuguese and Brazilian versus European Portuguese turned out to be substantially easier (at accuracies of 90% and 88%, respectively).

Classifiers that distinguish Dutch language varieties have also been developed. Trieschnigg et al. (2012) developed a classifier to discriminate between folktales written in Middle Dutch (the predecessor of modern Dutch, used in the Netherlands between 1200 and 1500) and 17th century Dutch, 20th century Frisian, and a number of 20th century Dutch dialects using the Dutch folktale database as a corpus. The performance of the classifier varied greatly per language variety: near-perfect to very good identification was achieved for some varieties (e.g. Frisian was identified with an F-score of 0.99; Liemers 0.88; Gronings 0.83), while classification was very difficult for other varieties (e.g. Overijssels at an F-score of 0.09; Waterlands 0.16; Drents 0.31). Tulkens et al. (2016) used corpora containing texts from mixed media (newspapers, Wikipedia, internet, social media) to build a Dutch language variety classifier based on provinces, and attained a relatively high score on some language varieties (up to 85% accuracy for Brabantian as spoken in the Belgian province of Antwerp), but they also report scores of 0% for six language varieties and a very low score on two others.

2.2 Features

While some exceptions exist (Tulkens et al., 2016), most of the current research in similar languages and language varieties use the same types of features, namely n-gram-based features. The results of the DSL shared task have shown that these approaches generally perform the best. However, scholars have argued that adding certain underused feature types could help improve the accuracy of state-of-the-art classifiers (Cimino et al., 2013). With the present study we investigate this claim by using two types of features in addition to word n-grams, namely text statistics (e.g. average word length, ratio of long/short words) and syntactic features (grammar-level features, e.g. PoS-tags).

Syntactic features have been used previously, though scarcely, in the context of language identification. Lui and Cook (2013) and Lui et al. (2014) used PoS n-grams as features for a classifier to make a distinction between English language varieties, while Zampieri et al. (2013) used PoS n-grams to classify Spanish language varieties. All three studies report that using POS n-grams leads to above-baseline results. This lends support to the notion that systematic differences between language varieties can be found using syntactic features.

The usage of text statistics for the identification of languages is even more uncommon compared to syntactic features. However, text statistics have been successfully used for similar research domains. One of these domains is native language identification (Jarvis et al., 2013; Cimino et al., 2013).

The successful implementation of text statistics features in this research domain implies that there

Language variant	Documents	Tokens
Netherlandic Dutch	77,430 (70%)	100,527,052 (68%)
Flemish Dutch	32,848 (30%)	47,888,260 (32%)
Total	110.278	148.415.312

Table 1: Document and token counts per language variety.

are systematic differences in stylistic choices between languages. A study by Windisch and Csink (2005) is one of the few studies using text statistics features for language identification. The authors found that these features can indeed be used for language identification. However, it should be noted that they studied dissimilar languages. The effectiveness of text statistics features for similar languages, or language variety identification remains an understudied subject.

2.3 Current work

The current study will explore lesser used techniques in the domain of language variety identification to see whether the current state-of-the art accuracy can be improved upon. This is done by using commonly used word n-grams together with the more uncommon lexical and syntactic features. Various approaches for combining these different feature types will be explored to investigate the added benefit of an ensemble classifier.

The current study focuses on the discrimination of Netherlandic Dutch (i.e. Dutch as spoken and written in the Netherlands) vs. Flemish Dutch (i.e. Dutch as spoken and written in the Dutch-speaking regions of Belgium). Speakers of Netherlandic Dutch and Flemish Dutch adhere to the same standard language, but, even so, linguists have stated that there are differences between Netherlandic and Flemish Dutch on every linguistic level, among which the lexical and syntactical level (De Caluwe, 2002). These differences tend to be subtle. Some examples of differences found between the two language varieties are word choice preference (e.g.*orange* in Netherlandic Dutch: *sinaasappel*, Flemish Dutch: *appelsien*), plural preference (e.g. *teachers* in Netherlandic Dutch: *leraren*, Flemish Dutch: *leraars*), and the order in which a particle and finite verb are preferably used (e.g. *I don't believe he* ***has come*** in Netherlandic Dutch: *Ik geloof niet dat hij* ***is gekomen***, Flemish Dutch: *Ik geloof niet dat hij* ***gekomen is***) (Schuurman et al., 2003).

Dutch language varieties have thus far remained a scarcely studied topic of research, although researchers have shown an interest in it. A limitation to the study of these varieties has always been the lack of available data (Zampieri et al., 2014). However, the recent introduction of the SUBTIEL corpus offers a usable corpus for such research. The feasibility of using this corpus is further explored in this work.

3 Method

3.1 Collection of the corpus

The SUBTIEL corpus contains over 500,000 subtitles in Dutch and English. These subtitles were produced by a professional studio operating in several countries, among which The Netherlands and Belgium. The procedure for these countries is mostly the same: a single translator provides the subtitles for a series episode or a movie. The main focus of the studio are movies and television shows, and to a smaller degree documentaries.

After filtering out the English subtitles and the Dutch subtitles without information on whether they were intended for Dutch or Flemish television, 110.278 documents remain; cf. Table 1. A document in this context is the subtitles for one movie, or one episode of a television show. For the subtitles used in this study, a distinction is made between subtitles that were shown on a Dutch or a Flemish television network. In comparison to similar work (Trieschnigg et al., 2012; Tulkens et al., 2016), the number of documents and tokens that is used in the current study is relatively large.

Using an automated mining tool, the subtitles in the corpus were scanned for a match in the Internet Movie Database (IMDb)[1], which provides additional information about the show or movie (e.g. genre, year, actors). The main interest was genre, since a vastly different genre distribution per language variety could have an impact on classification accuracy. An IMDb match was found for roughly half of the subtitles. The genre distribution for these matches did show minor dif-

[1] `http://www.imdb.com`

Genre	Netherlandic Dutch	Flemish Dutch
Drama	21.14%	25.11%
Comedy	14.51%	17.96%
Reality-TV	11.31%	5.63%
Crime	7.11%	9.40%
Action	5.52%	5.94%
Mystery	5.40%	4.95%
Documentary	5.88%	2.80%
Romance	5.56%	2.80%
Adventure	3.44%	3.33%
Family	2.87%	3.66%
Subtotal	83.15%	81.16%

Table 2: Distribution of the ten most frequent genres in the SUBTIEL corpus.

ferences between the language varieties, as can be seen in Table 2 For instance, the Netherlandic Dutch part of the corpus contained more subtitles for Reality-TV, Documentaries and Romance, while the Flemish Dutch part of the corpus contained more Drama and Comedy. Overall, the distribution of genres can be said to be reasonably similar.

Various types of information from the text were extracted as features to feed machine learning classifiers; cf. Table 3. Features were adopted based on previous work by Abbasi and Chen (2008) and Huang et al. (2010). The extracted features can be clustered into three groups: text statistics, syntactic features, and content-specific features. Text statistics features are based on counts at various levels (e.g. sentence/word length and word length distributions); syntactic features represent aspects of the syntactic patterns present in the data (e.g. the number of function words, punctuation and part-of-speech tag n-grams); content-specific features are any characters, character n-grams, words, or word n-grams that may be indicative of one particular language variant.

3.2 Classification methods

The five machine learning algorithms used in this study are AdaBoost with a decision tree core, C4.5, Naive Bayes, Random Forest Classifier, and Linear-kernel SVM. These types of algorithms have been used frequently for Language Identification tasks. SVM algorithms (Goutte et al., 2014; Malmasi and Dras, 2015; Jauhiainen et al., 2016) and Naive Bayes (King et al., 2014; Franco-Penya and Sanchez, 2016) are amongst the most popular algorithms. Decision tree approaches, which C4.5, AdaBoost, and Random Forest Classifier are examples of, have been used as well, but less frequently (Zampieri, 2013; Malmasi et al., 2016). The machine learning algorithms were deployed using the scikit-learn library (Pedregosa et al., 2011).

One of the challenges in the current study is to find an effective method of selecting the best combination of feature categories. One study on language variety classification has shown that an effective feature combination approach could increase classification accuracy (Malmasi et al., 2015). Three combination approaches are tested in the current study, namely the super-vector approach, two rule-based meta-classifiers, and one algorithm-based meta-classifier:

Super-vector All features, regardless of feature category, are merged into a single vector to predict the language variety.

Sum-rule meta-classifier The probabilistic outputs of the most accurate text statistics, syntactic, and content-specific classifier are summed, and the language variety with the highest sum is chosen.

Product-rule meta-classifier The product is calculated for the probabilistic outputs of the most accurate lexical, syntactic and content-specific classifier, and the language variety with the highest product is chosen.

Algorithm-based meta-classifier The probabilistic outputs of the most accurate lexical, syntactic and content-specific classifier are

Group	Category	Description	Number
Lexical	Average words per minute		1
	Average characters per minute		1
	Average word length		1
	Average sentence length in terms of words		1
	Average sentence length in terms of characters		1
	Type/token ratio	Ratio of different words to the total number of words	1
	Hapax legomena ratio	Ratio of once-occurring words to the total number of words	1
	Dis legomena ratio	Ratio of twice-occurring words to the total number of words	1
	Short words ratio	Words < 4 characters to the total number of words	1
	Long words ratio	Words > 6 characters to the total number of words	1
	Word-length distribution	Ratio of words in length of 1–20	20
Syntactic	Function words ratio	Ratio of function words (e.g. *dat*, *de*, *ik*) to the total number of words	1
	Descriptive words to nominal words ratio	Adjectives and adverbs to the total number of nouns	1
	Personal pronouns ratio	Ratio of personal pronouns (e.g. *ik*, *jou*, *mij*) to the total number of words	1
	Question words ratio	Proportion of wh-determiners, wh-pronouns, and wh-adverbs (e.g. *wie*, *wat*, *waar*) to the total number of words	1
	Question mark ratio	Proportion of question marks to the total number of end of sentence punctuation	1
	Exclamation mark ratio	Proportion of exclamation marks to the total number of end of sentence punctuation	1
	Part-of-speech tag n-grams	Part-of-speech tag n-grams (e.g. NP, VP)	Varies
Content-specific	Word n-grams	Bag-of-word n-grams (e.g. *lat*, *erg hoog*)	Varies

Table 3: Features adopted in our experiments.

used to train a higher level classifier, which is subsequently used to predict the language variety.

The algorithms tested as algorithm-based meta-classifier are the same algorithms that are used for the individual feature categories (AdaBoost, C4.5, Naive Bayes, Random Forest Classifier, and Linear SVM).

3.3 Processing and performance increases

Several preprocessing steps were undertaken. The goal for the content-specific classifier was to decrease the number of features, thus increasing processing speed, while retaining the most useful information. This was done by removing stop words, number strings and punctuation from the corpus: tokens that appear frequently, while carrying little meaning. Furthermore, words were normalized using lemmatization[2] to decrease the number of types for the content-specific features. Finally, words that did not appear more than 10 times in the corpus were removed.

To get the syntactic information necessary for the syntactic features, Pattern (Smedt and Daelemans, 2012) was applied to the texts, obtaining the part-of-speech tags. Part-of-speech tag n-grams that appeared less than 10 times in the corpus were removed.

After the frequency-based thresholding selection, another feature selection step was performed based on the chi-square weights of all features. Ranking the features and starting from the features with the largest weight, the subset of features was selected at which a saturation point was reached in performance on held-out data. No more than 10% of the features in the syntactic and content-free category turned out to be selected.

Besides steps to increase processing speed, steps to increase classification accuracy were also undertaken: hyperparameter optimization was applied to the algorithms. The optimal parameters were found by using 30-step Bayesian optimization on a random sample of 10% of the corpus.

[2]Lemmatization was performed with Frog, `https://languagemachines.github.io/frog/`

Method	Algorithm	# of features	Precision	Recall	F-score	Accuracy
Lexical only	AdaBoost	5	0.73	0.98	0.83	0.73
Syntactic only	AdaBoost	392	0.83	0.92	0.87	0.81
Content-specific only	Linear SVM	30,514	0.87	0.95	0.91	0.87
Lexical/Syntactic	AdaBoost	407	0.83	0.92	0.87	0.81
Lexical/Content-specific	AdaBoost	76.288	0.87	0.95	0.91	0.87
Syntactic/Content-specific	AdaBoost	76.325	0.87	0.95	0.91	0.86
Supervector	AdaBoost	76,325	0.86	0.94	0.90	0.86
Meta classifier (add)	-	-	0.87	**0.96**	0.91	0.87
Meta classifier (product)	-	-	0.87	**0.96**	0.91	0.87
Meta classifier (ML)	AdaBoost	6	**0.88**	**0.96**	**0.92**	**0.88**

Table 4: Classification performance.

4 Results

Table 4 lists the results obtained when classifying the Netherlandic Dutch and Flemish Dutch language varieties. Evaluation was done using 10-fold cross-validation and with precision, recall, F-score (with $\beta = 1$) and accuracy as metrics. Results range from a 73% accuracy score using lexical features only to 88% accuracy using an algorithm-based meta classifier. Thus, similar to Malmasi et al. (2015), the results of this study show that the best results are obtained when combining different types of features, using an algorithm-based meta-classifier.

AdaBoost appeared to be the most effective algorithm for most feature categories, except for the content-specific feature type, where the Linear-kernel SVM algorithm was the most accurate algorithm. This is in line with most DSL Shared Task entries, where the most common and accurate classifiers are SVM classifiers with content-specific features.

The recall values turn out to be particularly high, most of them above 0.95, while the precision scores are slightly lower: most of the classifiers obtained a score of around 0.85 for precision. This is further illustrated in Table 5, where a confusion matrix for the algorithm-based meta-classifier is shown: the classifier that obtained the highest performance.

The confusion matrix shows that Flemish Dutch documents were markedly harder to classify compared to Netherlandic Dutch documents. Nearly one third, 10,474 of the 32,848 Flemish documents, were incorrectly classified as Netherlandic Dutch, while a substantially smaller proportion of Netherlandic Dutch documents were incorrectly classified as Flemish Dutch (3208 out of 77,430). This may be partly explained by the fact that the number of Flemish Dutch documents is about half the number of Netherlandic Dutch documents in the SUBTIEL corpus.

Document language	Language variant	
	Flemish	Netherlandic
Flemish	22,374	10,474
Netherlandic	3208	74,222

Table 5: Confusion matrix for the algorithm-based meta-classifier.

4.1 Important features

The most important features per feature category are presented in Table 6. These features could be an indication of fundamental differences between the Netherlandic Dutch and Flemish Dutch language varieties and may therefore be useful from a linguistic perspective. The selection of feature importance is based on Random Forest Classification.

At the text statistics level, it can be observed that the ratio of words, especially shorter words, highlights important differences between Netherlandic Dutch and Flemish Dutch. There is a higher ratio of 1-, 2- and 5-letter words in the Flemish subtitles, while an average Netherlandic Dutch document contains more 3-letter words compared to Flemish Dutch documents, surprisingly. Additionally, sentences in Netherlandic Dutch subtitles contain more characters and words on average, and the ratio of words and characters per minute is higher in Netherlandic Dutch.

At the syntactic level, singular proper nouns (NNP) seem to be an important part-of-speech

Lexical	syntactic	Content-specific
Ratio of 1-letter words	NNP NN	nou
Ratio of 3-letter words	NNP PRP$	zandloper
Ratio of 5-letter words	NN FW	plots
Average amount of sentences in terms of words	, NNP	jij
Average amount of sentences in terms of characters	Personal pronouns ratio	hen
Ratio of 2-letter words	. PRP$	amuseren
Long words ratio	CD	orde
Words per minute	VB	vinden
Characters per minute	Function words ratio	lief helpen
Short words ratio	,	't

Table 6: Top 10 most important features per feature category.

category to discriminate Netherlandic Dutch from Flemish Dutch subtitles. Flemish subtitles have a higher ratio of sequences of singular proper nouns and singular nouns (NNP NN), singular proper nouns and possessive pronouns (NNP PRP$), and commas and singular proper nouns (, NNP). Furthermore, Flemish subtitles seem to contain a higher degree of singular nouns and foreign words (NN FW), periods and possessive pronouns (. PRP$), and commas (,), while Netherlandic Dutch subtitles contain more personal pronouns, cardinal numbers, and function words.

Some of the most important content-specific features indicate typical lexical differences between language varieties. For instance, *nou* has been previously noted to be a word that is not used as much in Flemish as compared to Netherlandic Dutch,[3] and *plots* is noted to be a word used more in Flemish.[4] No such categorical status is known for the other important content-specific features, although *amuseren* and *lief helpen* may arguably be associated more with Flemish Dutch. *Zandloper*, *jij*, *hen*, and *orde* also appeared more frequently in Flemish subtitles compared to Netherlandic Dutch, while *vinden* and *'t* appeared more in Netherlandic Dutch subtitles. The relative importance of some of these features in the current task could be due to hidden artifacts of the corpus.

5 Conclusion and future work

In this paper we presented language identification experiments carried out with five machine learning techniques (AdaBoost, C4.5, Naive Bayes, Random Forest Classifier, and Linear SVM), and three feature categories (text statistics, syntactic features, and content-specific features) focusing on the Netherlandic and Flemish variants of Dutch. Subtitles collected in the SUBTIEL corpus were used to train and test the classifiers on. With the exception of a few studies (Lui and Cook, 2013; Lui et al., 2014; Windisch and Csink, 2005; Zampieri et al., 2013), text statistics and syntactic features have rarely been explored in language identification tasks. Additionally, there are not many classification studies focusing on Dutch language varieties, exceptions being Trieschnigg et al. (2012) and Tulkens et al. (2016).

The highest accuracy score was obtained when using a meta-classifier approach with a machine-learning algorithm, AdaBoost. In this approach the probabilistic scores obtained from classifiers trained exclusively on text statistics features, syntactic features, and content-free classifiers respectively were used as input for training a higher-level classifier. This result is in agreement with the findings of Malmasi et al. (2015), where the best results were also obtained using a meta classifier. This result suggests that a meta-classifier approach is a viable approach to language (variety) identification, and also supports the claim by Cimino et al. (2013) that underused feature types such as text statistics and syntactic features could improve classification accuracy. Furthermore, most of the classifiers performed best using an AdaBoost algorithm with decision tree core.

The accuracy, precision, recall and F-measure scores obtained with the algorithm-based meta-classifier are substantially higher than scores obtained with previous Dutch language variety clas-

[3] http://www.taaltelefoon.be/standaardtaal-verschillen-tussen-belgie-en-nederland

[4] http://taaladvies.net/taal/advies/vraag/665/plotsklaps_eensklaps_plots_plotseling/

sifiers. Trieschnigg et al. (2012) obtained an F-score of 0.80 versus the F-score of 0.92 in this study, and Tulkens et al. (2016) achieved an average accuracy of around 15% versus 88% in this study. Furthermore, the results seem to be on par with state-of-the-art methods: Zampieri et al. (2016) obtained accuracy scores between 74% and 90% in the binary classification of newspaper texts in variants of Portuguese, and Malmasi et al. (2015) obtained accuracy scores between 76% and 94% for binary classification of Arabic language varieties.

However, it is important to note that direct comparison between the current work and previous language variety identification studies is likely to be misleading. In this study, the classification of language varieties was based on the country the subtitle was developed for. It was not based on the country the subtitle writer was originally from, since this information was not known. Furthermore, Zampieri et al. (2016) and Malmasi et al. (2015) have shown that classification accuracy could be markedly different depending on how closely related the language varieties are, Lui and Cook (2013) have shown that different corpora could result in different accuracy scores, and the amount of language varieties that a classifier discriminates between has an effect on the accuracy as well. Thus, the difference between this study and the studies of Trieschnigg et al. (2012) and Tulkens et al. (2016) could be a matter of different corpora, corpus size, and the fact that the classifier in this study discriminated between two language varieties while the classifiers of Trieschnigg et al. (2012) and Tulkens et al. (2016) between sixteen and ten varieties, respectively.

Therefore, it would be interesting to see how the current approach competes against other approaches using the same corpus. When competing in such a task, it would be interesting to test whether the performance of the current approach could be further increased, for instance by including character-level features in the lexical and content-specific feature categories, since all the features in the current work reside at the word-level. Windisch and Csink (2005) have shown that character-level lexical features (word endings, character ratios, consonant congregations) are useful features for the classification of different languages, and character n-grams are one of the most popular features for language classification (Zampieri, 2013). Furthermore, partial replication of the current study could be interesting with modifications to the current corpus and algorithms. Accuracy scores could change if the Netherlandic Dutch and Flemish Dutch data are balanced and if proper names are removed from the corpus (Zampieri et al., 2015). There are also different types of meta-classifiers (e.g. a voting-based meta-classifier) and algorithms (e.g. XGBoost, Multilayer Perceptron) that were not tested in the current study and that might improve classification accuracy, which is worth further exploration.

The ranked list of most useful features found in this work could be a basis for future linguistic research on differences between Netherlandic Dutch (as spoken mainly in the Netherlands) and Flemish Dutch (as spoken mainly in Flanders). The findings for the lexical features suggest a difference in text difficulty between Netherlandic Dutch and Flemish Dutch texts: Flemish subtitles contain a higher ratio of short words, shorter sentences and generally less text. We would like to stress that these results could be do to differences in the SUBTIEL corpus. More research would be necessary to investigate whether such a stylistic difference between Netherlandic Dutch and Flemish Dutch exists outside of the SUBTIEL corpus.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29.

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistic profiling based on general-purpose features and native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–215, Stroudsburg, PA, June. Association for Computational Linguistics.

Johan De Caluwe. 2002. In *Taalvariatie en taalbeleid : bijdragen aan het taalbeleid in Nederland en Vlaanderen*. Garant.

Hector-Hugo Franco-Penya and Liliana Mamani Sanchez. 2016. Tuning Bayes Baseline for dialect detection. In Preslav Nakov, Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Shervin Malmasi, editors, *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties*

and Dialects, pages 227–234, Stroudsburg, PA, December. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 139–145, Stroudsburg, PA, August. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1800–1807, Paris, France, May. European Language Resources Association.

Chunneng Huang, Tianjun Fu, and Hsinchun Chen. 2010. Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, 61(5):891–906.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of The 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Stroudsburg, PA, June. Association for Computational Linguistics.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. HeLI, a word-based backoff method for language identification. In Preslav Nakov, Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Shervin Malmasi, editors, *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 153–162, Stroudsburg, PA, December. Association for Computational Linguistics.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Stroudsburg, PA, August. Association for Computational Linguistics.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In Sarvnaz Karimi and Karin Verspoor, editors, *Proceedings of the Australasian Language Technology Association Workshop 2013*, pages 5–15, Stroudsburg, PA, December. Association for Computational Linguistics.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 2014. Exploring methods and resources for discriminating similar languages. In Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 129–138, Stroudsburg, PA, August. Association for Computational Linguistics.

Wolfgang Maier and Carlos Gómez-Rodriguez. 2014. Language variety identification in Spanish tweets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 25–35, Stroudsburg, PA, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409, Stroudsburg, PA, June. Association for Computational Linguistics.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In Kôiti Hasida and Ayu Purwarianti, editors, *Proceedings of the Fourteenth International Conference of the Pacific Association for Computational Linguistics*, pages 35–53, Singapore, May. Springer.

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 Preslav Nakov, Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Shervin Malmasi, editors, *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Stroudsburg, PA, December. Association for Computational Linguistics.

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton Van der Wouden. 2003. CGN, an annotated corpus of spoken Dutch. In *Proceedings of 4th International Workshop on Language Resources and Evaluation*, pages 340–347.

Tom de Smedt and Walter Daelemans. 2012. Pattern for Python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In Martin Kay, editor, *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2619–2634, Stroudsburg, PA, May. Association for Computational Linguistics.

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska de Jong, and Theo Meder. 2012. An exploration of language identification techniques for the Dutch folktale database. In Petya Osenova, Stelios Piperidis, Milena Slavcheva, and Cristina Vertan, editors, *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage*, pages 47–51, May.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4130–4136, Paris, France, May. European Language Resources Association.

Gergely Windisch and László Csink. 2005. Language identification using global statistics of natural languages. In *Proceedings of the Second Romanian-Hungarian Joint Symposium on Applied Computational Intelligence*, pages 243–255, May.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of the Confrence sur le Traitement Automatique des Langues Naturelles 2013*, pages 580–587, Stroudsburg, PA, June. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Stroudsburg, PA, August. Association for Computational Linguistics.

Marcos Gebre Zampieri, Binyam Gebrekidan, Hernani Costa, and Josef van Genabith. 2015. Comparing approaches to the identification of similar languages. In Preslav Nakov, Marcos Zampieri, Petya Osenova, Liling Tan, Cristina Vertan, Nikola Ljubešić, and Jörg Tiedemann, editors, *Proceedings of the Second Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 66–72, Stroudsburg, PA, September. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Octavia-Maria Şulea, and Liviu P. Dinu. 2016. A computational approach to the study of Portuguese newspapers published in Macau. In Larry Birnbaum, Octavian Popescu, and Carlo Strapparava, editors, *Proceedings of the Workshop on Natural Language Processing Meets Journalism*, pages 47–51, Stroudsburg, PA, July. Association for Computational Linguistics.

Marcos Zampieri. 2013. Using bag-of-words to distinguish similar languages: How efficient are they? In *Proceedings of the Fourteenth Symposium on Computational Intelligence and Informatics*, pages 37–41, New York, NY, November. Institute of Electrical and Electronics Engineers.