An Exploration of Language Identification Techniques in the Dutch Folktale Database


Published in
Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage (LREC 2012)
2012

document version
Peer reviewed version
document license
CC BY

Link to publication in KNAW Research Portal

citation for published version (APA)

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the KNAW public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the KNAW public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
pure@knaw.nl

Download date: 14. May. 2024
An Exploration of Language Identification Techniques for the Dutch Folktale Database

Dolf Trieschnigg 1, Djoerd Hiemstra 1, Mariët Theune 1, Franciska de Jong 1, Theo Meder 2

1 University of Twente, Enschede, the Netherlands
2 Meertens Institute, Amsterdam, the Netherlands
{d.trieschnigg,d.hiemstra,m.theune,f.m.g.dejong}@utwente.nl, theo.meder@meertens.knaw.nl

Abstract
The Dutch Folktale Database contains fairy tales, traditional legends, urban legends, and jokes written in a large variety and combination of languages including (Middle and 17th century) Dutch, Frisian and a number of Dutch dialects. In this work we compare a number of approaches to automatic language identification for this collection. We show that in comparison to typical language identification tasks, classification performance for highly similar languages with little training data is low. The studied dataset consisting of over 39,000 documents in 16 languages and dialects is available on request for followup research.

1. Introduction
Since 1994 the Meertens Institute 1 in Amsterdam has been developing the Dutch folktale database, a large collection of folktales in primarily Dutch, Frisian, 17th century and Middle Dutch and a large variety of Dutch dialects (Meder, 2010). It does not only include fairy tales and traditional legends, but also riddles, jokes, contemporary legends and personal narratives. The material has been collected in the 19th, 20th and 21st centuries, and consists of stories from various periods, including the Middle Ages and the Renaissance. The database has an archival and a research function. It preserves an important part of the oral cultural heritage of the Netherlands and can be used for historical and contemporary comparative folk narrative studies. An online version has been available since 2004 2 and currently contains over 41,000 entries.

A rich amount of metadata has been assigned manually to the documents, including language, keywords, proper names and a summary (in standard Dutch). This metadata is very useful for retrieval and analysis, but its manual assignment is a slow and expensive process. As a result, the folktales database grows at a slow rate. The goal of the FACT (Folktales as Classifiable Text) 3 research project is to study methods to automatically annotate and classify folktales. Ideally, these techniques should aid editors of the folktales database and speed-up the annotation process. Language identification is the first challenge being addressed in the FACT project.

In this paper, we compare a number of automatic approaches to language identification for this collection. Based on the performance of these approaches we suggest directions for future work. The Dutch folktale database poses three challenges for automatic language identification. First, the folktales are written in a large number of similar languages. A total of 196 unique language combinations is present in the metadata; 92 unique (unmixed) language names are used 4 . For most of these languages no official spelling is available; the way words are spelled depends on the annotator who transcribed the oral narrative. As a result, documents in the same language may use a different spelling. For our experiments we have used a selection of 16 languages. Second, the language distribution in the collection is skewed: most of the documents are in Frisian and Standard (or modern) Dutch, but there is a long tail of smaller sets of documents in other languages. Consequently, for many languages only little training data is available to train a classifier. Third, documents in the collection can be multilingual. Most of the documents are monolingual, but some contain fragments in a different language. The length of these fragments ranges from a single passage or sentence to multiple paragraphs.

The contributions of this work are twofold. First, we present an analysis of multiple language identification methods on a challenging collection. Second, we make this collection available to the research community.

The overview of this paper is as follows. In Section 2 we briefly discuss related work. In Section 3 we describe the collection in more detail and outline the experimental setup. In Section 4 the results of the different classification methods are discussed followed by a discussion and conclusion in Section 5.

2. Related work
Early work on language learnability dates back to the 1960s (Gold, 1967). Since the 1990s language detection or language identification has become a well-studied natural language processing problem (Baldwin and Lui, 2010). For clean datasets, with only few and clearly separable languages, language identification is considered a solved problem (McNamee, 2005).

Recent research indicates, however, that language identification still poses challenging problems (Hughes et al., 2006), including: supporting minority languages, such as the dialects encountered in our collection; open class language identification, in such a way that a classifier is capable of indicating that no language could be accurately determined; support for multilingual documents; and classification at a finer level than the document level. Xia et al. (2009) and Baldwin and Lui (2010) also argue that language identification has not been solved for collections con-

1 http://www.meertens.knaw.nl
2 http://www.verhalenbank.nl (in Dutch only)
3 http://www.elab-oralculture.nl/FACT
4 Sometimes caused by an inconsistent naming convention
taining large numbers of languages. In this work we will focus on the capability of existing classifiers to deal with minority and very similar languages.

A large array of methods has been developed for tackling the problem of language identification: categorization based on n-grams (Cavnar and Trenkle, 1994), words or stopwords (Dasamsh, 1995; Johnson, 1993), part-of-speech tags (Grefenstette, 1995), syntactic structure (Lins and Gonçalves, 2004), systems based on markov models (Dunning, 1994), SVMs and string kernel methods (Kruengkrai et al., 2005), and information theoretic similarity measures (Martins and Silva, 2005). An extensive overview of techniques is outside the scope of this paper. A more comprehensive overview can be found in Hughes et al. (2006) and Baldwin and Lui (2010). We limit our experiments to the method by Cavnar and Trenkle (1994) and a number of variations based on n-grams and words motivated by positive experimental results of (Baldwin and Lui, 2010).

3. Experimental setup

In the following subsections we describe the collection, investigated classification methods, and evaluation metrics in detail.

3.1. The collection

The complete folktale database consists of over 41,000 documents. After filtering out documents with offensive content (sexual, racist, lese-majesty, etcetera) and copyrighted materials, 39,510 documents remain. From this collection we put all documents with a mixed language where at least one of the languages is Standard Dutch into a single language group labeled “Standard Dutch mixed”. Documents in a language with fewer than 50 documents in that language in the collection are removed. This results in a collection of 39,003 documents in 16 different languages. Table 1 lists the 16 languages and their document frequencies. Note that the number of documents per language is strongly skewed: 79% of the collection is written in Frisian or Standard Dutch. The remaining 21% of the documents is distributed over the remaining fourteen languages. Also note that in comparison to previous work by Baldwin and Lui (2010), which uses collections between 1500 and 5000 documents, the collection is relatively large.

3.2. Classification methods

As a baseline classification method, we used the TextCat implementation of the algorithm described by Cavnar and Trenkle (1994). The algorithm creates an n-gram profile for each language and performs classification by comparing each of the n-gram profiles to the n-gram profile of the text to classify. An out-of-place distance measure is used to compare the order of n-grams in the profile and the text. Following the methods investigated by Baldwin and Lui (2010) we used a number of classification methods based on nearest neighbour (NN) and nearest prototype (NP) in combination with the cosine similarity metric. All tested classification methods use a supervised learning approach: classifications are based on a training set of manually labeled examples. The difference between NN and NP methods is the way the examples are stored. In the NP case, the examples of the same class are aggregated into a prototype, a single model representing the class. The prototype is constructed by summing the vectors of the examples. In the NN case, the examples are stored separately. During classification the class(es) of the nearest example(s) is/are returned. In our case we use the class of the first nearest neighbour (or prototype).

The documents are represented by vectors of the unit of analysis, containing the count of that unit. In the case of words, each unique word encountered in the collection forms one dimension of the vector. We use six different units of analysis: overlapping character n-grams of size 1 to 4, a combined representation of n-grams of length 1 to 4, and words (uninterrupted sequences of letters). The text is lowercased and punctuation is removed before features are extracted. The overlapping character n-grams are extracted by sliding a window of n characters over the text one character at a time. In case of the combined n-gram representation, this process is repeated four times (for n=1 to n=4). To reduce the complexity, we experiment with reducing the vector to a selection of 100, 500 and 1000 features. The selection of features is based on the most frequently used features per language appearing the training set. To be more precise: from each language the most frequent feature is taken until the desired number of features is reached. In our experiments we follow the approach described by Baldwin and Lui (2010). Alternatively, we could have used information gain to select the most informative features. We will consider this in future work.

3.3. Evaluation method

We evaluated the different approaches by means of stratified 10-fold cross-validation: the collection was split into

<table>
<thead>
<tr>
<th>Language</th>
<th>Doc. count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frisian</td>
<td>17,347</td>
</tr>
<tr>
<td>Standard Dutch</td>
<td>13,632</td>
</tr>
<tr>
<td>17th century Dutch</td>
<td>2,361</td>
</tr>
<tr>
<td>Standard Dutch mixed</td>
<td>1,538</td>
</tr>
<tr>
<td>Flemish</td>
<td>882</td>
</tr>
<tr>
<td>Gronings¹</td>
<td>854</td>
</tr>
<tr>
<td>Noord-Brabants¹</td>
<td>677</td>
</tr>
<tr>
<td>Middle Dutch</td>
<td>656</td>
</tr>
<tr>
<td>Liemers¹</td>
<td>328</td>
</tr>
<tr>
<td>Waterlands¹</td>
<td>153</td>
</tr>
<tr>
<td>Drents¹</td>
<td>150</td>
</tr>
<tr>
<td>Gendts¹</td>
<td>116</td>
</tr>
<tr>
<td>English</td>
<td>97</td>
</tr>
<tr>
<td>Overijssels¹</td>
<td>80</td>
</tr>
<tr>
<td>Zeeuws¹</td>
<td>68</td>
</tr>
<tr>
<td>Dordts¹</td>
<td>64</td>
</tr>
<tr>
<td><strong>Total (16 languages)</strong></td>
<td><strong>39,003</strong></td>
</tr>
</tbody>
</table>

¹ Dutch dialects

Table 1: Collection statistics
4. Results

4.1. TextCat baseline

Table 2 lists the classification performance of TextCat for the 16 languages in the collection. The contingency table in table 3 provides further information about the classification errors made. Its rows list the actual classes where its columns indicate the predicted classes indicated by the system. For example, the second row and first column indicates that 6 documents in Standard Dutch were incorrectly classified by TextCat as Frisian.

We can make the following observations. First, the classification performance of the largest language class (Frisian) is very good. The recall is very high (0.98) at almost perfect precision (0.999). Second, the classification performance of old Dutch languages (17th century Dutch and Middle Dutch) is also good (F-measure larger than 0.96). These languages can be distinguished well from modern Dutch and dialects. Third, the classification performance of the dialects is mixed. Some (Liemers, Gronings) perform relatively well, others (Waterlands, Overijssels) perform poorly. Still, the highest F-measure (0.88) does not come close to typical performance scores, which range between 0.91 and 0.99 for the EuroGOV collection (Baldwin and Lui, 2010). Most of the dialects are mistaken for Standard Dutch and vice versa. Gronings shows strong overlap with Drents (both northern dialects); Zeeuws is frequently mistaken for Noord-Brabant, but not the other way around (both southern dialects). Fourth, it is striking that classification of English documents is so poor. Table 3 indicates that Standard Dutch and Standard Dutch mixed is frequently mistaken for English. One possible explanation is that English words or expressions are frequently borrowed in Dutch. It could also indicate that the annotation in the collection is inconsistent: the (Dutch) document contains an English expression but has been classified as Standard Dutch instead of Standard Dutch mixed.

The micro average performance score (see Table 2) indicates a reasonable classification performance of TextCat, but this value has been strongly influenced by the strong performance on the largest language class. The macro averages illustrate that for many smaller languages classification performance is low. Figure 1 shows a scatter plot of the amount of training data available for a language and its classification score.

4.2. Variations of cosine distance

Table 4 summarises the classification performance of a number of variations on language identification systems. TextCat can be viewed as a variation of a nearest prototype method (in terms of F-measure). All the cosine variants of the nearest prototype method performed poorly. Still the highest F-measure (0.88) does not come close to typical performance scores, which range between 0.91 and 0.99 for the EuroGOV collection (Baldwin and Lui, 2010). Most of the dialects are mistaken for Standard Dutch and vice versa. Gronings shows strong overlap with Drents (both northern dialects); Zeeuws is frequently mistaken for Noord-Brabant, but not the other way around (both southern dialects). Fourth, it is striking that classification of English documents is so poor. Table 3 indicates that Standard Dutch and Standard Dutch mixed is frequently mistaken for English. One possible explanation is that English words or expressions are frequently borrowed in Dutch. It could also indicate that the annotation in the collection is inconsistent: the (Dutch) document contains an English expression but has been classified as Standard Dutch instead of Standard Dutch mixed.

The micro average performance score (see Table 2) indicates a reasonable classification performance of TextCat, but this value has been strongly influenced by the strong performance on the largest language class. The macro averages illustrate that for many smaller languages classification performance is low. Figure 1 shows a scatter plot of the amount of training data available for a language and its classification score.

4.2. Variations of cosine distance

Table 4 summarises the classification performance of a number of variations on language identification systems. TextCat can be viewed as a variation of a nearest prototype method (in terms of F-measure). All the cosine variants of the nearest prototype method performed poorly. Still the highest F-measure (0.88) does not come close to typical performance scores, which range between 0.91 and 0.99 for the EuroGOV collection (Baldwin and Lui, 2010). Most of the dialects are mistaken for Standard Dutch and vice versa. Gronings shows strong overlap with Drents (both northern dialects); Zeeuws is frequently mistaken for Noord-Brabant, but not the other way around (both southern dialects). Fourth, it is striking that classification of English documents is so poor. Table 3 indicates that Standard Dutch and Standard Dutch mixed is frequently mistaken for English. One possible explanation is that English words or expressions are frequently borrowed in Dutch. It could also indicate that the annotation in the collection is inconsistent: the (Dutch) document contains an English expression but has been classified as Standard Dutch instead of Standard Dutch mixed.

The micro average performance score (see Table 2) indicates a reasonable classification performance of TextCat, but this value has been strongly influenced by the strong performance on the largest language class. The macro averages illustrate that for many smaller languages classification performance is low. Figure 1 shows a scatter plot of the amount of training data available for a language and its classification score.
Table 3: Contingency matrix for TextCat

<table>
<thead>
<tr>
<th>Actual ↓</th>
<th>Frisian</th>
<th>Standard Dutch</th>
<th>17th century Dutch</th>
<th>Standard Dutch mixed</th>
<th>Flemish</th>
<th>Gronings</th>
<th>Noord-Brabants</th>
<th>Middle Dutch</th>
<th>Liemers</th>
<th>Waterlands</th>
<th>Drents</th>
<th>Gendts</th>
<th>English</th>
<th>Overijssels</th>
<th>Zeeuws</th>
<th>Dordts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frisian</td>
<td>16928</td>
<td>106</td>
<td>5</td>
<td>31</td>
<td>27</td>
<td>78</td>
<td>3</td>
<td>5</td>
<td>48</td>
<td>53</td>
<td>7</td>
<td>48</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Standard Dutch</td>
<td>6</td>
<td>8630</td>
<td>13</td>
<td>366</td>
<td>2065</td>
<td>12</td>
<td>446</td>
<td>6</td>
<td>20</td>
<td>1028</td>
<td>195</td>
<td>1</td>
<td>554</td>
<td>156</td>
<td>4</td>
<td>130</td>
</tr>
<tr>
<td>17th century Dutch</td>
<td>21</td>
<td>2308</td>
<td>3</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td>12</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Dutch mixed</td>
<td>1</td>
<td>695</td>
<td>9</td>
<td>135</td>
<td>194</td>
<td>15</td>
<td>165</td>
<td>7</td>
<td>4</td>
<td>131</td>
<td>32</td>
<td>92</td>
<td>48</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Flemish</td>
<td>124</td>
<td>1</td>
<td>1</td>
<td>714</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>24</td>
<td>2</td>
<td>20</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gronings</td>
<td>45</td>
<td>3</td>
<td>3</td>
<td>670</td>
<td>19</td>
<td>1</td>
<td>14</td>
<td>84</td>
<td>1</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noord-Brabants</td>
<td>118</td>
<td>1</td>
<td>6</td>
<td>63</td>
<td>8</td>
<td>378</td>
<td>2</td>
<td>5</td>
<td>18</td>
<td>25</td>
<td>3</td>
<td>47</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle Dutch</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>639</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liemers</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>298</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waterlands</td>
<td>15</td>
<td>9</td>
<td>2</td>
<td>126</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drents</td>
<td>4</td>
<td>1</td>
<td>28</td>
<td>106</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gendts</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>10</td>
<td>3</td>
<td>16</td>
<td>65</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overijssels</td>
<td>21</td>
<td>7</td>
<td>8</td>
<td>18</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zeeuws</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>23</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dordts</td>
<td>11</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The nearest neighbour cosine variants perform similar or better than TextCat in terms of micro and macro F-measure. It should be noted, however, that these nearest neighbour approaches are far more expensive in terms of processing time and required storage than the method implemented by TextCat. Second, the cosine variants perform better with longer representations (longer n-gram windows or words) and with more features. Using all features performs best, but the selection of 1000 features closely approximates the scores based on all features. Figure 2 illustrates the differ-
ence between TextCat and the (NN) Cosine distance with word features: Cosine performs better on all languages, except Middle and 17th century Dutch, and Gronings.

5. Conclusions and future work
In this work we have investigated a number of language identification methods on a new and large collection of folktales in a variety and mix of languages. Compared to other nearest prototype methods, the approach based on mixed n-grams proposed by Cavnar and Trenkle (1994) performs well. The results showed that a nearest neighbour approach using longer and more features performs even better.

Compared to other language identification tasks carried out by Baldwin and Lui (2010), the classification results stay behind. Baldwin and Lui (2010) report a maximum macro F-measure of 0.729 for a skewed collection containing 67 languages. With similar methods, we achieve only 0.630, for a collection with fewer languages. These results indicate that this collection indeed poses a challenge for language identification. The collection therefore is a valuable resource for future language identification research. The collection is available on request (users are required to sign a license agreement).

An important note has to be made on the consistency of the language annotations in the collections. The folktales in the database have been gathered and annotated (in a free text field) by more than 50 people. It is an open question whether these editors have used the same method for labelling the language of a document; some might have annotated a document with Standard Dutch, where another would have labeled it as a mix of Standard Dutch and another language. This might explain why the automatic methods cannot discriminate between these classes.

Our future work will focus on the following aspects of language identification. First, we intend to focus on multilingual document detection. Almost 10% of the documents in the complete collection contain multiple languages. Therefore, it would be useful to detect languages at the sentence level. Second, it would be useful to assign a level of certainty to the detected language. In the work described in this paper we view the task as a closed classification problem with a fixed number of languages. Especially for the long tail of documents in minority languages it would be useful to indicate if no known language was confidently determined. Third, since the language identification system is intended to be used in a semi-automatic setting, it is useful to have a mechanism to present proof for the detected language. Especially when the annotator has no in-depth knowledge of the different languages this would be useful. This could be achieved, for example, by showing sentences from the suggested language(s) similar to the sentence under classification. Fourth and finally, since classification performance is still relatively low, we intend to investigate how contextual information can be used to improve classification performance. In the line of recent work from Carter et al. (2013), who improved the language identification of Twitter messages by incorporating classification features based on for example language of the blogger and language of the document linked to, we could introduce additional features for this particular domain. One can think of features based on the date, source, and place of narrative of the folktale. Or a feature based on the geographical locations encountered in the text. In addition, it might be possible to incorporate knowledge from dialect lexicons to improve classification.

6. Acknowledgements
This work has been carried out within the Folktales as Classifiable Texts (FACT) project, which is part of the CATCH programme funded by the Netherlands Organisation for Scientific Research (NWO).

7. References