German Dialect Identification and Mapping for Preservation and Recovery

Aynalem Tesfaye Misganaw, Sabine Roller
Universität Siegen
aynalem.misganaw@uni-siegen.de, sabine.roller@uni-siegen.de

Abstract

Many linguistic projects which focus on dialects do collection of audio data, analysis, and linguistic interpretation on the data. The outcomes of such projects are good language resources because dialects are among less-resources languages as most of them are oral traditions. Our project *Dialektatlas Mittleres Westdeutschland* (DMW)[1] focuses on the study of German language varieties through collection of audio data of words and phrases which are selected by linguistic experts based on the linguistic significance of the words (and phrases) to distinguish dialects among each other. We used a total of 7,814 audio snippets of the words and phrases of eight different dialects from middle west Germany. We employed a multilabel classification approach to address the problem of dialect mapping using Support Vector Machine (SVM) algorithm. The experimental result showed a promising accuracy of 87%.

Keywords: language identification, dialects, less-resourced languages

1. Introduction

For widely used languages like English, French, German etc. the problem of language identification (LI) is addressed because language resources are available in significant amount. In contrast, finding language resources for dialects is a challenging task since they are among less-resourced languages, which makes it to be a bottleneck when it comes to employing language technologies.

In the DMW project, systematic data collection is performed both on conducting a survey to select speakers and interviewing them. The speakers are from different regions in middle west Germany which includes North Rhine-Westphalia, parts of Lower Saxony and Rhineland-Palatinate. The interview questions are designed in such a way that the various linguistic aspects like vocabulary (lexicon), word structure and word formation (morphology), sound structure (phonology) and sentence formation (syntax) could be analyzed, evaluated, and interpreted for identifying the non-standard way of speaking i.e., dialects.

The collected data contains, among other metadata descriptions, the audio snippets, their transcriptions in IPA (International Phonetic Alphabet) notation, the words in focus (in standard German), and the region which the speakers represent. The audio snippets contain the spoken utterances of the selected words, phrases, and sentences. In this paper, acoustic features are extracted from each audio snippets into a csv format.

The use of stop words, n-gram, Machine Learning (ML) and hybrid approaches are commonly used method of LI (Truica et al., 2015). All these methods require the use of a significant size of language resources. For resource-rich languages, the process of identifying a language, using either of the methods, is relatively easy as they have writing standards from which rules could easily be extracted. However, in addition to the lack of standard in writing and the scarcity of written resources, identifying the nuances of dialects is a challenging task.

The distinction among dialect is so fine unlike the most widely used languages where a list of frequently used words could be used to distinguish them. Nowadays, the web is a good source for linguistic resources making this work to have a great deal of significance in **crowd corpus collection** which is a key input for corpus-based research. In addition, for researchers in the field of linguistics, it will have a benefit of **mapping a particular dialect with the region** it is spoken.

Many Natural Language Processing (NLP) either assume the language they are dealing with or use LI in their pipeline before trying to solve the main problem. This work will benefit those researchers who are struggling to support the preservation of less-resourced language.

This paper is organized as follows. Related work is briefly reviewed in Section 2. The dataset description and the process of identifying the parameters are described in detail in section 3 and 4 respectively. Section 5 discusses the method employed in identifying and mapping dialects. Results and discussions are explained in section 6. Finally, we provide conclusion and recommendations for future works in section 7.

2. Related Work

There is a significant number of work on LI with the aim of developing a system which is able to recognize and infer a language under question (Jauhiainen et al., 2019) . In the survey they conducted, Jauhiainen et al.

[1] https://www.dmw-projekt.de/

showed that LI could be applied to any form of language; text, speech, and sign language; digital or otherwise. Although notable progress has been achieved for resource-rich languages in the last couple of decades, less-resourced languages and dialects do not yet benefit from the state-of-the art technologies (Chittaragi and Koolagudi, 2019).

Among the main methods used for language recognition are methods based on stop words, character n-grams, machine learning and hybrid (Truica et al., 2015). In addition, the commonly used features in solving the problem are spectral and prosodic features (Chittaragi and Koolagudi, 2019)(Bartz et al., 2017) (Cai et al., 2019), transcripts of speech data (Ramesh et al., 2021)(Malmasi and Zampieri, 2017), and written text (Truica et al., 2015).

Scannell (2007), Jauhiainen et al. (2020) and Jurgens et al. (2017) have applied LI with the aim of corpus construction for endangered languages. Web services like automatic translation need to first recognize the language before translating the content into a target language (Lui and Baldwin, 2012). Thus, LI is critical in most language processing problems where its low performance affects the whole pipeline as it propagates (Jauhiainen et al., 2019).

3. Description of the Data

The data used is from the DMW project where people representing different places in Middle West Germany are selected and interviewed. The data used in this work represents eight geographical locations (Wenkerort[2]) each representing different dialect varieties.

The data collection is done by directly interviewing dialect speakers. The interview is based on a questionnaire which contains 800 questions, a sample of which is shown in Table 1. The interview is conducted by asking the speakers a question, and by showing them a video or image and let the speakers describe it in their dialect. The questions are designed in a way aimed at getting the translation of names of objects, animals, and activities in a dialect.

The complex tasks of data collection and the subsequent preprocessing are done by employees of the project in four different partner Universities, among which 12 are responsible for the field work of interviewing dialect speakers. The interview takes about 3 to 5 hours and sometimes it takes three different sessions to get the complete interview.

Although the data contains the linguistic features like vocabulary (lexis), word structure and word formation (morphology), sound structure (phonology) and sentence formation (syntax), this work focused on the use of lexicons for identifying and mapping the dialects. The total number of audio snippets used in the study is 7814 representing eight distinct dialects from middle northwest Germany.

Question	English Translation	Note
Welches Tier ist das auf dem Bild?	Which animal is on the picture?	A picture of a goat is shown to the interviewee
Wie lautet die Mehrzahl von Ziege?	What is the plural of goat?	
Worauf kann man reiten?	What can you ride on?	
Was macht die Frau in dem Video gerade?	What is the woman in the video doing right now?	A video is shown for the interviewee.

Table 1 Sample Questions

During the interviews, a tool - *SpeechRecorder* (Draxler and Jänsch, 2004) - is used. This tool enabled to store only part of the interview which is relevant.

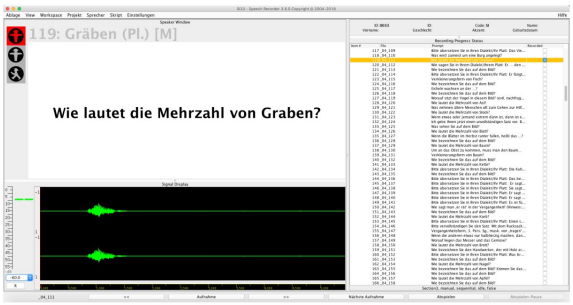

Figure 1: SpeechRecorder Software

For example, for the question text shown in Figure 1 where the question displayed means "*What is the plural of ditch?*", the speaker might utter other words before or after he speaks the answer for the question. However, using the tool the interviewer can record only the relevant part. In addition to the Speechrecorder tool, using a web-based interface further data cleaning is done in which the audio files are further cropped so that the audio exactly matches required answer.

[2] Named after the famous German linguistic researcher Georg Wenker.

Each speaker and each question are uniquely identified. The combination of these IDs is used to label the audio data. The region the speakers represent is also identified by unique ID which is used to label the dialect varieties.

Table 2 shows the distribution of audio files per dialects and the number of speakers used in each dialect region. Thus, the number of audio snippets used in this study ranges from 297 to 1303 per dialect. As presented in Table 1, except for the two dialect places, *Glehn* and *Homberg*, all the other six dialect have two speakers each.

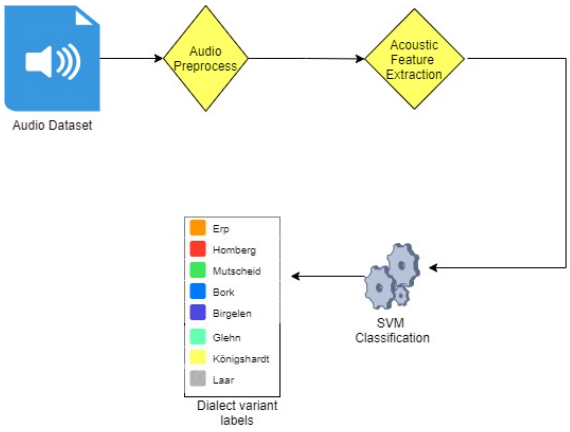

Figure 2 Block diagram for feature extraction and dialect identification

Although there are some audio snippets with longer time, the average length of the audio snippets ranges from 1sec to 5 sec.

Dialect Region	No. of Audio Snippets	No. of speakers	Total Length in min:sec
Birgelen	1301	2	19:05
Bork	1301	2	19:51
Erp	898	2	16:18
Glehn	297	1	04:14
Homberg	303	1	04:05
Königshardt	1303	2	19:49
Laar	1140	2	17:29
Mutscheid	1271	2	25:49

Table 2 Table showing the size of audio data and number of speakers for each dialect region

4. Parameter Identification/Feature extraction

Although the focus of the interview is collection of dialectal data, it is conducted in standard German. As a result, the audio files sometimes contain utterances which are unrelated to the question. This makes data cleaning an inevitable task. Thus, the audio recordings are first cropped, shown in the preprocessing step of Figure 2, to match only the utterances in dialect related to the question at hand.

After the audio preprocessing, the next process as shown in Figure 2 is acoustic feature extraction. The spectral and temporal acoustic features are extracted from the audios snippets using librosa (McFee et al., 2021). The features used in this study are shown in Table 3.

These acoustic features are used to extract audio properties like the pitch, energy, rise and fall of the frequency, and melody of the speaker.

MFCC are series of values which collectively make up an MFC (*Mel-frequency Cepstrum).* These values could range from 1 to 39, which could be generated using the audio feature extraction and manipulation module of the librosa package (McFee et al., 2021). This feature is important in that it helps identify and

Spectral Features
Chroma feature
root-mean-square (RMS)
spectral centroid
spectral bandwidth
spectral rolloff
zero crossing rate
Mel frequency cepstral coefficients (MFCC)

Table 3 List of spectral features

represent how human sounds are produced by vocal tract. The shape of the vocal tract like tongue, teeth, lips, nasal cavity, etc. determines the sound generated by humans. Correctly determining and representing this shape enables the correct representation of phonemes in the sound generated. This shows that although audio data contains utterance of words and/or phrases, the acoustic features extraction makes it possible that phoneme level properties are captured and represented in numeric format. Accordingly, in this study 20 MFCC features of each audio snippets are used.

5. Classification of Dialect Varieties

The problem of dialect identification in this work is dealt as document classification using SVM classification algorithm where the labels for

documents correspond to the dialect variety and the acoustic features as documents.

The dialect varieties used in this paper are eight (*Erp, Homberg, Mutscheid, Bork, Birgelen, Glehn, Königshardt, and Laar*), where one audio snippet corresponds to one dialect. The dialect variants are named after the Wenker place the speakers represent.

The dataset shows that the number of distinct classes are the same as the number of dialects at hand. In our case, the class labels for the classification problem are eight. Accordingly, we have employed a multilabel classification method using SVM in which the dialect variants, i.e., the labels for the classification problem are converted to multi-class labels using the scikit-learn label converter.

The dataset shows that there is subtle difference among the dialects. SVM is chosen as it is capable of drawing boundary mid-way between closest points of any two classes in a dataset.

$$S = \begin{bmatrix} f_{11} f_{12} \ldots f_{13} \cdots f_{1z} \\ f_{21} f_{22} \ldots f_{23} \cdots f_{2z} \\ \ldots \\ f_{n1} f_{n2} \ldots f_{n3} \cdots f_{nz} \end{bmatrix}$$

Notation 5.1

Notation 5.2 shows the acoustic features f_{iz} of the dataset of sample S of size n. In our case the number of features denoted by z, i.e., the total number of features is 26 (20 MFCC and the other six spectral features shown in Table 3).

The classification output of any given sample s_i shown in the form of Notation 5.1 is represented as a set of C values $\{c_1, c_2, c_3, \ldots c_m\}$ where c_i are labels for the given dialect region for s_i.

$$C = \{c_1, c_2, c_3, \ldots c_m\}$$

Notation 5.2

where c_i are elements in class C of size m, i.e., eight.

6. Results and discussions

The experiment is done in two phases. In the first phase only data of a single speaker per dialect is used. In the second phase, data from the second speaker is added to the dataset.

In addition, although the dataset contains 20 MFCC features, we experimented to see the difference in the accuracy of the classifier using different number of MFCC. Hence, as the number of MFCC feature used increases, the model showed improvement which is illustrated in Table 4.

The model is trained with 67% of the dataset and evaluated with the rest. Splitting the data is done in a way that the model does not do unintended classification having only one speaker as test set. The train and test are randomly selected to avoid a test set containing only one speaker. Thus, using the score metrics, we achieved a promising result of 91%.

No. of MFCC	Score
11	0.77
12	0.79
14	0.81
15	0.82
16	0.84
17	0.85
18	0.85
19	0.86
20	0.87

Table 4 List of scores based on the number of MFCC used

7. Conclusion and Future Work

This work assumes that a particular word or phrase is uttered uniquely in all the eight dialect regions. However, there are words which are pronounced the same in different dialect regions. So, we would like to recommend considering the problem as a multi-label and multi-class problem, i.e., a particular row in the dataset can assume more than one dialect variant as its label.

There are many more dialect variants in Germany, beyond the data used in this research. If the identification of German dialects includes the other varieties, it would increase the contribution to the less-resourced languages and thereby to NLP technologies in general.

Acknowledgment

The audio snippets are taken from the interviews conducted in the DMW project. For building and evaluating the model, the OMNI computing cluster of the Universität Siegen is used.

This work is Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 262513311 – SFB 1187

References

Bartz, C., Herold, T., Yang, H., Meinel, C., 2017. Language identification using deep convolutional recurrent neural networks, in: Internationalconference neural information processing. pp. 880–889.

Cai, W., Cai, D., Huang, S., Li, M., 2019. Utterance-level End-to-end Language Identification Using Attention-based CNN-BLSTM, in:

ICASSP20192019IEEEInternationalConference Acoustics, SpeechSignalProcessing(ICASSP). pp. 5991–5995.
Chittaragi, N.B., Koolagudi, S.G., 2019. Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms. Language Resources and Evaluation 1–33.
Draxler, C., Jänsch, K., 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software, in: Proc. LREC. Lisbon, pp. 559–562.
Jauhiainen, H., Jauhiainen, T., Lindén, K., 2020. Building Web Corpora for Minority Languages, in: Proceedings 12th WebCorpusWorkshop. European Language Resources Association, Marseille, France, pp. 23–32.
Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., Lindén, K., 2019. Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research 65, 675–782.
Jurgens, D., Tsvetkov, Y., Jurafsky, D., 2017. Incorporating dialectal variability for socially equitable language identification, in: Proceedings 55th Annual MeetingAssociation Computational Linguistics (Volume 2: Short Papers). pp. 51–57.
Lui, M., Baldwin, T., 2012. langid. py: An off-the-shelf language identification tool, in: Proceedings ACL 2012 system demonstrations. pp. 25–30.
Malmasi, S., Zampieri, M., 2017. German Dialect Identification in Interview Transcriptions, in: Proceedings Fourth WorkshopNLP Similar Languages, VarietiesDialects(VarDial). Association for Computational Linguistics, Valencia, Spain, pp. 164–169.
McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Ellis, D., Mason, J., Battenberg, E., Seyfarth, S., Yamamoto, R., viktorandreevichmorozov, Choi, K., Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.-R., Friesch, P., Weiss, A., Vollrath, M., Kim, T., Thassilo, 2021. librosa/librosa: 0.8.1rc2.
Ramesh, G., Vayyavuru, V., Rama, M.K.S., 2021. Attention-Based Phonetic Convolutional Recurrent Neural Networks for Language Identification, in: 2021NationalConference Communications (NCC). pp. 1–6.
Scannell, K.P. (Ed.), 2007. The Crúbadán Project: Corpus building for under-resourced languages.
Truica, C.-O., Velcin, J., Boicea, A., 2015. Automatic language identification for romance languages using stop words and diacritics, in: 201517thInternationalSymposium SymbolicNumericAlgorithms Scientific Computing (SYNASC). pp. 243–246.