AUTOMATIC KURDISH DIALECTS IDENTIFICATION

Hossein Hassani[1,2] and Dzejla Medjedovic[3]

[1]Department of Computer Science and Engineering,
University of Kurdistan Hewlêr, Erbil, Kurdistan Region, Iraq
[2]Department of Computer Science, Sarajevo School of Science and Technology,
Sarajevo, Bosnia and Herzegovina
hosseinh@ukh.edu.krd
hossein.hassani@stu.ssst.edu.ba
[3]Department of Computer Science, Sarajevo School of Science and Technology,
Sarajevo, Bosnia and Herzegovina
dzejla.medjedovic@ssst.edu.ba

ABSTRACT

Automatic dialect identification is a necessary Language Technology for processing multi-dialect languages in which the dialects are linguistically far from each other. Particularly, this becomes crucial where the dialects are mutually unintelligible. Therefore, to perform computational activities on these languages, the system needs to identify the dialect that is the subject of the process. Kurdish language encompasses various dialects. It is written using several different scripts. The language lacks of a standard orthography. This situation makes the Kurdish dialectal identification more interesting and required, both form the research and from the application perspectives. In this research, we have applied a classification method, based on supervised machine learning, to identify the dialects of the Kurdish texts. The research has focused on two widely spoken and most dominant Kurdish dialects, namely, Kurmanji and Sorani. The approach could be applied to the other Kurdish dialects as well. The method is also applicable to the languages which are similar to Kurdish in their dialectal diversity and differences.

KEYWORDS

Dialect identification, NLP, Kurdish language, Kurmanji, Sorani

1. INTRODUCTION

Dialectology has not received a considerable attention in Computational Linguistics (CL) and Natural Language Processing (NLP). However, it has been part of sociolinguistics and linguistics in general for a long time. For example, dialectal classification study for English [1] and other European languages started several decades ago. It has produced different results such as preparation of "Atlas Linguarum Europae" and "European linguistic map" [2]. Although the situation for the languages such as English and German that have been the focal subject of CL, does not show any radical change with regard to computational dialectology, it seems to be

David C. Wyld et al. (Eds) : ITCS, JSE, SIP, ARIA, NLP - 2016
pp. 61–78, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60307

slightly shifting for other languages. To illustrate, recent works express some interests in computational dialectology for languages such as Arabic and Chinese [3, 4].

Several reasons could have caused the mentioned "ignorance". First, a language such as English, which has received a high degree of attention from computational perspective, has a standard (or two main standards: American and British) format. Neither the speakers nor the readers of several dialects of this language have serious difficulties in understanding each other. Second, the language is written with one script and follows a standard orthography. Third, the dialects are constructed on a common architecture of the language with insignificant variation in their structures.

Although other reasons could be counted, for this research, these three can justify the necessity for the development of the methods and the tools, as part of the Language Technology (LT), for a multi-dialect language such as Kurdish for which one or more of the mentioned reasons are not applicable.

The LT follows the same principles that the technology generally follows in almost all cases. That is, its development and improvement are rooted in different needs such as usual day-to-day, sociological, and business related requirements. By the same analogy, it is also developed and evolved based on scientific, experimental, and laboratory demands that usually happen in the research activities. This might justify why the dominant languages in NLP have not attracted such attention, while other languages such as Chinese and Arabic have noticeably been of interest for dialectal studies from computational perspective.

The same analogy could be used for Kurdish, which is the language of Kurds. Kurdish is spoken in divergent dialects. The speaking population of this language is uncertain. There is a discrepancy in the reports of population ranging from 19 million [5] to 28 million [6]. In fact, the Kurds' lives as well as their language have been extremely affected by their political and geopolitical situation. Indeed, they have remained marginalized geographically, politically, and economically, during the last two centuries [7]. Consequently, the negligence of their language has been propagated to many other fields. Therefore, it is not of any surprise to see that the Kurdish computational study is not by any mean an established sector of CL and NLP.

Diversity of Kurdish dialects, grammatical distances, vocabulary differences, and mutual unintelligibility are some main factors in guiding most of the scarce CL and NLP research and development into two main dialects of Kurdish, namely Kurmanji and Sorani. In fact, in most cases for the reasons that we will address in section 3, these activities have focused on Sorani alone. This research has also focused on the two stated dialects for the same reasons. Regardless of the number of the dialects, dialect identification is a requirement in Kurdish CL and NLP. The reason is, when one wants to perform a computational process on Kurdish text, such as Machine Translation (MT), sentiment analysis, Part of Speech Tagging (POST) or a similar process, one cannot proceed without being aware of the dialect of the context. We will discuss this case in more detail in the following sections.

The rest of the article has been organized in five sections. The first section discusses the dialects, languages and their relations. The second section provides an overview of Kurdish, its dialects, scripts, grammar, orthography, and some linguistic aspects of the language. The third section explains the related works and the methodology of the research. The forth section discusses the

experiments, the results and their analysis, and some found issues. The Conclusion section summarizes the findings and suggests some areas that need more study in the future.

2. DIALECTS AND LANGUAGES

Although the definition of the language seems to be clear, when one wants to distinguish between "dialects" and "languages", the borderlines seem to be blurred [8]. Linguists have different opinion about dialects and languages. Nevertheless, they mostly agree on referring to two sets of criteria, one social and the other political, based on which one can distinguish whether what is spoken in a specific community is a dialect or a language [9].

Despite the fact that the similarity in the vocabulary, the pronunciation, the grammar, and the usage are important parameters in making distinction between dialects and languages, the central concept of this distinction is suggested to be the concept of mutual intelligibility [9]. That is, if two dialects are mutually unintelligible, they are considered two different languages, otherwise, two dialects of the same language. However, there are dialects that are mutually intelligible which are considered languages and there are others which are mutually unintelligible yet considered dialects of a language [9].

In the case of Kurdish dialects, the definition is arguable from some linguists' point of view [10]. This situation makes research activities on Kurdish dialects rather challenging. This is because although the basis of CL is common among dialects and languages, it might significantly affect the research approach and methodology according to whether you tackle the problem area *interlingually* or *intralingually*. The following section provides more information on this subject.

3. KURDISH LANGUAGE

Kurdistan is the homeland of the Kurds. This is a region located across Iran, Iraq, Turkey, and Syria. The term Kurdistan has been used to name a province or to call a wider region, both in Iran and Iraq. For some political reasons, the case is different in Turkey and Syria. Kurdish people are sometimes called "a nation without state" [11]. However, the Kurds have been in the middle of many battles over the past centuries, hence, their geopolitics situation has always been a matter of concern to the world's policy.

Surprisingly, this situation has not benefited them as much as it has the other communities with the similar status. For example, regardless of the utmost cruelty that happened, several countries received benefits out of both World Wars, while the Kurdish population never received such benefits until recent times. As a result, this situation has affected the Kurdish usage and its popularity as well. It seems that the circumstances are going to be different since the Iraqi Kurdistan region has started to have its regional government under the new federal Iraq.

In the following sections a background on Kurdish, its dialects, scripts, orthography, and its current situation with regard to CL and NLP will be discussed.

3.1. Overview

Kurdish is the name given to a number of distinct dialects of a language spoken in the geographical area touching on Iran, Iraq, Turkey, and Syria. However, Kurds have lived in other

countries such as Armenia, Lebanon, Egypt, and some other countries since several hundred years ago. They also have large diaspora communities in some European countries and North America.

There are some opinions about the Kurdish root that state that the Kurds have come from different origins, that they have changed their language, and their first language was rather different from the current one [12]. However, those who believe in this theory do not make this clear that how people from different origins have spoken an unknown language and why they have changed it. More accurate figure on this has been given by [13].

Kurdish studies, though not very popular, has an almost a century of history. McCarus provides an informative background on Kurdish studies. Although his work dates back to the 1960s, it still can be seen as a major resource about the Kurdish studies [14]. A very recent finding based on a different approach "to try to prove with inter-disciplinary scientific methods explained, that indigenous aborigine forefathers of Kurds (speakers of the 'Kurdish Complex') existed already B.C.E. and had a prehistory in their ancestral homeland (mainly outside and Northwest of Iran of today)" [15]. However, research on Kurdish has been biased for different reasons.

Kurdish language includes different dialects. Dialect diversity is an important characteristic of Kurdish. Kurdish is written using four different scripts. The popularity of the scripts differ according to the geographical and geopolitical situations. There is no consensus among the Kurdish linguists upon the number of letters in the Kurdish alphabet. Latin script uses a single character while Persian/Arabic and Yekgirtû in some cases use two characters for one letter. The Persian/Arabic script is even more complex with its right-to-left and concatenated writing style.

Kurdish is spoken in different dialects, which are not following the same grammar. The level of the differences vary for every pair of dialects. I addition, an important feature of current Kurdish is the lack of a standard orthography [16].

The above brief overview shows the complexity of Kurdish from different perspectives. This complexity, particularly, affects the language computation, which in turn makes hindrance in front of studying Kurdish in the context of CL and NLP. It also makes the development of LT for this language rather challenging. Below we discuss these issues in a more detail.

3.2. Dialects

As it was mentioned, Kurdish is a multi-dialect language. Since the 1960s several major scholars, including Westerns and Kurds have published influential research outcomes about Kurdish and its dialects [16, 19–21]. But neither the nomenclature of these dialects have been standardized nor there is a solid agreement on their relation to the language [13, 22, 23].

In a recent research, Haig and Öpengin identify the Kurdish dialects as Northern Kurdish (Kurmanji), Central Kurdish (Sorani), Southern Kurdish, Gorani, and Zazaki. For each one of these dialects they mention the main sub-dialects [18]. The populations that speak different dialects of the language differ significantly. The majority of Kurmanji speakers are located in different countries, such as Turkey, Syria, Iraq, Iran, Armenia, and Lebanon, just to name the main lands. The second popular dialect is Sorani, which is mainly spoken among Kurds in Iran and Iraq. Zazaki is spoken in Turkey. Gorani is primarily spoken in Iran and Iraq [16, 21]. In

addition, as a result of long-term conflicts in the region, Kurds also have a large diaspora community in different western countries, where almost all Kurdish dialects are spoken to different extents.

It is worth mentioning that as Leezenberg describes, the reason that we stick with the internationally accepted categorization of the language is to keep the harmonious environment of the scholarly research [24], while we are aware and actually prefer, at least due to the local usage, to use Hawrami, or Hawramani, instead of Gorani, in most of the situations. In fact, as we have observed, sometimes, the term Gorani, is rather restricted, if it is not considered as unknown at all, in most of Sorani and Hawrami dominated areas.

3.3. Alphabet and Scripts

Kurdish is written using four different scripts, which are modified Persian/Arabic, Latin, Yekgirtû(unified), and Cyrillic [25]. The popularity of the scripts differ according to the geographical and geopolitical situations. There is no consensus among the Kurdish linguists upon the number of letters in the Kurdish alphabet. The main reason for the disagreement seems to be mainly on the phonetic aspects (and to a great extent acoustic features) rather than lexical aspects, though clearly these two affect each other. For example, Bedir_Xan and Lescot suggested 31 letters in their Latin script proposal for Kurmanji, arguing that the Kurdish did not have a separate sound to distinguish between 'خ' and 'غ', hence in their Latin script they used the letter "x" for both sounds [26]. However, these two sounds are written with two different letters in Persian/Arabic and Yekgirtû scripts. As a result, some sounds are lost if an utterance is captured using Latin script. In order to address this issue, the current Latin script has been augmented to capture the mentioned sounds [25]. Nonetheless, as an advantage, Latin script uses a single character while Persian/Arabic and Yekgirtû, in some cases, (e.g., 'وو' in Persian/Arabic and 'sh' in Yekgirtû for 'û' and 'ş' in Latin, respectively) use two characters for one letter. Although Yekgirtû is phonetically more complete (it includes 37 "letters"), its double character representation for a single phoneme makes it computationally more difficult. The Persian/Arabic script is even more complex with its right-to-left and concatenated writing style.

Latin script, mainly, is used for writing in Kurmanji dialect. But this is not applied for Kurdish communities in Armenia and former Soviet countries, whom they use Cyrillic script. Furthermore, until recently, the Kurmanji community of the Iraqi Kurdisatn, was mostly using modified Persian/Arabic script. For Sorani, the main script is modified Persian/Arabic. Zazaki, mainly, is written in Latin. Gorani (Hawrami) is, mainly, written in modified Persian/Arabic. We stressed on word "mainly", because there are considerable exceptions in the usage of these scripts, particularly Latin and modified Persian/Arabic. The former is used in Turkey, because the Kurd community is already familiar with the script through Turkish. In Iran, Syria, and Iraq, the dominance is with the Persian/Arabic script. The reason is obvious. Persian is the national and formal script in Iran, and Arabic has been the national and formal script for Iraq and Syria. Generally speaking, Persian/Arabic script has a longer history in writing Kurdish, while Latin script was suggested and introduced by Mir Celadet Bedir-Xan around 1930s [27].

However, in the recent years the situation has been changing. That is, the Latin script is growing in the usage and is becoming more popular in the areas that it was not before. But it is not the same for Persian/Arabic or Cyrillic script. That is, the Persian/Arabic has been continuing to be the dominant script in the areas that it was and the Cyrillic script has been restricted to the

communities in Armenia and the former Soviet countries. As an example for the latter, Persian/Arabic script is the official script in the Iraqi Kurdistan Region, though the usage of Latin script is growing, particularly, by different Kurdish media.

3.4. Grammar and Orthography

Despite having the same root, Kurdish dialects grammatically differ from each other. The differences are vary in terms of grammatical features and the level that they differ [16–18]. In some cases the grammatical differences are trivial, while in some others they are considerable. We show this with two samples.

As the first sample, Sorani speakers do not apply gender differentiation, while Kurmanji applies gender. To be more precise, there are restricted sub-dialects, which is spoken by a small community of a population of less than a few thousand people. The authors, in their research, have recently come across a small Sorani speaking community, where the gender is used, to just differentiate between male and female human-being. Indeed, Hassanpour has already addressed the issue of genders and its usage in some sub-dialects of Sorani [16], which does not seem that is in use by the current speakers anymore.

In another case, the authors learned that similar situation is true for one of the sub-dialects of Laki, which is called Jafar-aabadi. Laki, itself, is a sub-dialect of Southern Kurdish, which in general does not include genders. However, the authors were told, by an informant about the dialect, that Jafar-aabadi speakers use gender, not only for human-being but also for other subjects. As an example, in the small community where this dialect is spoken, the Moon is masculine. The reason that is mentioned for this assignment is that the Moon dares to come out at nights, while the Sun is feminine, because it comes out during the day. Further investigation on this case is a linguistic endeavor. However, authors are following the case as it is related to their other areas of research in CL.

The second sample is the negation, where in Kurmanji one says "**ne** li nêzîkê", which means "it is not close", while in Sorani it is said a "le nêzîk **niye**" (The negations were shown in bold). These examples show the difficulty of dealing with Kurdish as "a language" and not as "a group of languages".

Kurdish has different issues from an orthographic point of view. First and foremost, there is no standard orthography for Kurdish. Hassanpour gives a brief history of how an orthography was suggested in Iraq based on the Arabic language alphabet and the challenges that it faced during the 1920s [16]. Finding the reasons for why a language that is spoken by a massive population had not have its own orthography, which in turn sparkles other questions such as why the written Kurdish only has a history of no more than a few centuries, is not an easy task.

For example, some sources, orally, have talked to one of the authors about the correspondence between the Arab army and Kurds defending Banah (a Kurdish city located in the Kurdistan province in Iran) around the 670s. These sources based their "story" upon information that they have received from a descendant family, whom were involved in that correspondence. Although the author could not ascertain the case at this stage, it was worth mentioning for further investigations. Even if this story is not true, it is still not clear what kind of alphabet and orthography have been used by the Kurds at the time. Although this is basically a matter that is

related to linguists and historians, if we find some reliable answers, we will share it with the interested researchers.

3.5. Kurdish and Language Technology

In this research we use the "computationally-enabled" as a technical term to distinguish the languages that are enabled with the minimum tools of Language Technology, which in turn allows those languages and their products, whether in written or spoken format, to be processed by computers. Although there are frameworks for this accounting and assessment such as BLARK (The Basic Language Resource Kit) [28], this is beyond the scope of this article. We just mention the case in the capacity of the current article.

Despite having a large speaking population, there is no or limited computational research with regard to Kurdish. Indeed, a simple search on the Internet regarding computational activities on the language provides no more than a few results, which either they are at the preliminary stages of the study, or they cover very specific areas such as text to speech concepts, to discuss some limited corpus, or to provide some comparison between the dominant Kurdish dialects.

Moreover, even this small amount of research mostly covers one dialect of Kurdish, which is Sorani. Therefore, currently Kurdish cannot be considered as a computationally-enabled language. Although Hassani and Kareem discuss the case with regard to assistive technologies for Kurdish [29] and there are also other appreciable attempts by some scholars which have taken place [30], the overall figure has neither been progressing significantly nor is promising. In spite of the fact that there are some evidence showing a slight growth in the interest in this area (for example, see [31]), to become a computationally-enabled language, Kurdish needs extensive scholarship and professional efforts.

3.6. Current Situation

As a consequence of the establishment of the Iraqi Kurdistan Regional Government, Kurdish has become one of the two official languages. This has been declared under Article 4 of the Iraqi Constitution [32]. Neither this article nor other part of the constitution specifies a particular dialect of Kurdish in this regard. Similar approach has been followed in the Draft Constitution of Kurdistan Region, which has been approved by the Parliament of Kurdistan Region [33, 34] (The Draft Constitution of Kurdistan Region is in Arabic and Kurdish; a translation into English can be found here [35]). The document can become official if it is approved in a referendum, which has not been held yet.

As a result of the above steps, Kurdish has become the main teaching medium for the entire pre-university education. Even though there are some exceptions for the private schools, which might use some foreign languages such as English or French, in these schools too learning Kurdish is an obligatory educational element. Furthermore, the language is used to a varying extent in most of the universities in the region as well.

However, the decision on making a dialect official depends on the population who speak the dialect in the specific area/governorate. For instance, Kurmanji is the official dialect for communication and education (up to the end of high school) in Duhok governorate of the Iraqi Kurdistan Region, while in the other two governorates Sorani dialect plays the same role.

There is no precise demographic report accessible to show the population who speak different dialects, but the figure can be loosely extracted from the population who live in different governorates. A report on the Iraqi Kurdistan Population Forecast for 2009-2020 period shows that about 26% of the Iraqi Kurds are currently living in the Kurmanji dominant areas [36]. These demographic facts vary significantly in different countries where Kurds live. As an illustration, the figures for the language shows that Kurmanji is spoken by around 20 million Kurds [37], Sorani is spoken by around 7 million [38], and other dialects or sub-dialects (for example, Hawrami, Kalhori, Feyli, and others) are spoken by around 3 million Kurds [39].

Nevertheless, alongside promoting linguistic diversity and rights in general, "the [Iraqi] Kurdistan Regional Government's policy is to promote the two main dialects [of Kurdish] in the education system and the media" [40]. Consequently, the majority of satellite TVs in the Iraqi Kurdistan Region has at least news programs broadcasting in both dialects, either at the same session or as the separate sessions. In fact, some TV channels display news tickers (crawlers) in both dialects and sometimes in both Latin and Persian-Arabic scripts (e.g., Kurdistan TV, Rudaw, Kurdsat, NRT, and KNN). The websites of some of these TV channels are also provided in both dialects [41-43]. But, perhaps because the majority of Iraqi Kurds speak Sorani and this dialect has a long rich historical and literature background in the Iraqi Kurdistan Region, the de facto dialect of the conversations and the documents of the Iraqi Kurdistan Region is Sorani [44].

However, in spite of the emerging usage of the language, both regionally [45] and worldwide, Kurdish is not yet official in other countries where Kurds live. The reason behind this brief explanation is not to highlight political issues and motivations but to outline the fact that Kurdish might play a more significant role in the coming years. Without considering other parts of Kurdistan, being the official language of the Iraqi Kurdistan region only, suggests that Kurdish Computational Linguistics needs considerable attention.

Nevertheless, as the context of languages has changed tremendously because of the emergence and rapid spread of information technology, to become a well-known language in the world, Kurdish needs to be studied in light of the paradigm of CL and NLP. In fact, Kurdish needs to be understood not only by other people throughout the world but also among the Kurds themselves who speak different dialects that are not mutually intelligible.

Furthermore, Kurdish has a low visibility among the Internet users. Also currently there are no machine translator, no optical character recognizer, no commercialized text to speech, and no speech to text systems available for the language. In fact, crucial issues such as lack of grammatical and orthographical standards for the language, would affect any attempt towards the development of such utilities for the *intralanguage/interlanguage* purposes.

In summary, many obstacles stand in the way of advancing the preparation of Kurdish in order to be computationally processed. Moreover, working on Kurdish from a Computational Linguistics and Natural Language Processing perspectives would require some fundamental elements. For example, developing a corpus, as a core element that is required for many aspects of CL and NLP such as machine translation, dictionary preparation, text classification, discourse analysis, and text summarization, needs a substantial amount of time, budget, and effort. This becomes more challenging if one thinks about having a specific corpus for each special domain of the language study and processing. The challenge can grow if this corpus should be kept up-to-date and accessible to different users.

Equally important, for some reasons, which are beyond the scope of this article, written literature for the Kurdish language does not have a diverse and lengthy background. For some dialects/sub-dialects such as Kalhori and Hawrami, the case might be even more serious.

Moreover, Kurdish CL and NLP have not yet been established as academic disciplines. A quick survey on the available websites of universities, which are located in the Kurdish speaking areas in Iraq, Turkey, Iran, Syria, and Armenia, shows no fact that these subjects have been taken seriously except in one case, University of Kurdistan – Sanandaj, which one can find some valuable studies, though focusing mainly on one of the Kurdish dialects [46]. Indeed, current academic research on the Kurdish language in terms of CL and NLP is neither established nor seems to be promising as a scientific research area.

In this research we have tried to take one step towards an important issue with regard to Kurdish CL and NLP, which is automatic dialect identification. The following chapter explains the methodology of the research.

4. METHODOLOGY

Dialectology has been one of the research areas in traditional linguistics for almost as long as linguistics has been recognized as an independent field of science. However, the same is not the case in the Computational Linguistics context, at least for the dominant languages in the field such as English, German, and French. Therefore, when one is interested in computational dialectology, soon finds that the major works in this area have been carried out for some languages which, in computing sense, are not very popular.

To illustrate, Kessler has provided a method for computational dialectology in Irish Gaelic [47]. Similarly, Nerbonne and Heeringa have worked on Dutch [48]. They have computationally compared and classified 104 Dutch dialects. This type of dialectology assumes that the dialects under the investigation are mutually intelligible. In most of the cases, the focus is on the *phone* differences or slight changes that happen in the language morphology, from one dialect to the other.

In a different context, Tang and Heuven have performed a series of thorough experiments on some Chinese dialects in which they have provided some methods for these dialects classification [4]. In this latter case, intelligibility among the dialects is the main concern of the research.

Another research has been carried out in order to identify the Arabic dialects, which has resulted in suggesting an annotator that is used to annotate the Arabic texts according to their identified dialects [49].

Text classification is a well-studied area in Natural Language Processing, yet it still is a very demanding research subject [50–52]. Most of the text classification methods concentrate on the context classification. Different methods are used in text classification, most of which are based on Machine Learning techniques [53].

In the current research, we adapted a text classification method in order to classify Kurdish texts into the dialects that the texts are written in. We have targeted two main Kurdish dialects: Kurmanji and Sorani. The adapted method was applied in several steps, namely, data collection,

transliteration, and weighting list creation. Finally, the outcomes were tested in order to investigate the accuracy of the dialect identification. These steps will be explained in the following sections.

4.1. Transliteration

As it was mentioned, Kurmanji texts are, mainly, written in Latin script, while for Sorani texts the main script is Persian/Arabic. Persian/Arabic script has a longer history, while Latin script was suggested and introduced by Mir Celadet Bedir-Xan around the 1930s [27]. However, in both cases exceptions exist. That is, one can find texts in Kurmanji that have been written in Persian/Arabic script and texts in Sorani that have been written in Latin script. Again, as it was mentioned, currently no standard orthography exists for either dialect.

For this research, we collected the texts from different Kurdish media. In addition, it was decided to use the Latin script as the base for the dictionaries, the training set, and for the test data as well. But, because the Sorani texts were mainly written in Persian/Arabic script, the texts had to be transliterated into Latin script. In order to do so, we have developed a tool (a transliterator) in Python that transliterates the texts which are written in Persian/Arabic script into Latin script.
The main challenge of this transliteration process is the lack of a standard orthography in writing Kurdish. This case was discussed in section 3.4. Also there are ambiguous cases that an automatic transliterator is not able to produce what one might be able to produce by manual transliteration.

Our Python transliterator uses three compact Python dictionaries in order to cover the three different cases, which occur in the Kurdish writing using Persian/Arabic script. The first Python dictionary includes digits and single characters that can be transliterated into a single equivalent Latin character. For example, ‘ک’ and ‘ك’ both would be transliterated to 'k'. The second Python dictionary includes double characters, which have been concatenated using a special connector. It also handles the situations where the code and the shape of the concatenated characters are changed due to the participation in the concatenation. In this situation, in some cases, a double character must be transliterated to one character, and in some others, to two equivalent characters. For example, ‘ئا’ would be transliterated to 'a', while 'پ' would be transliterated to 'p'. The third Python dictionary is used for the situations where a character is concatenated to its predecessor or successor using two concatenation connectors on its both sides or it includes a postfix space such as 'ب' and 'په', which would be transliterated to 'b' and 'pe' respectively.

The transliterator was tested and tuned to cover all cases which are known to be special. The mentioned Python dictionaries were ordered and tuned manually. However, lack of standard orthography causes that one cannot expect that the result of the transliteration to be correct in all cases. Nevertheless, the result of the transliteration was tested manually in different situations in order to make sure that the transliterator produces the correct results when one compares the results against the original texts.

Figure 1 shows a sample in Kurmanj, which has been written using Persian/Arabic script.

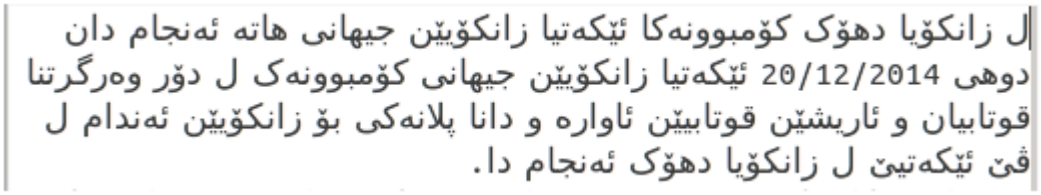

Figure 1. A sample text of Kurmanji in Persian-Arabic script

Figure 2 shows the result of the transliteration of the text of Figure 1 using the developed transliterator.

```
l zankoya dhok kombûneka êketya zankoîên cîhany hate
encam dan
dwhy 20/12/2014 êketya zankoîên cîhany kombûnek l dor
wergrtna qutabian u arîşên qutabiên aware u dana
planeky  bû zankoîên endam l vê êketîê l zankoya dhok
encam da.
```

Figure 2. The transliterated text of Figure 1

4.2. Weighting List Creation

In this research, we have used an adaptation of Support Vector Machines (SVM) [54], [55]. For the training and test phases, we collected data from different resources available on the Internet. For this purpose, we used the websites of several Kurdish media. The fundamental reason for this approach was because we decided to restrict our study to the most contemporary concepts that were widely understandable by the target dialects speakers.

At this step, which can be interpreted as the training phase, the classifier reads the texts and "sanitizes" the text to remove non-alphabet characters from the text, using regular expressions. It then extracts the vocabulary of the text and inserts them into a weighting matrix. We decided to include only the words with the length of at least two characters in this weighting matrix. Obviously, duplication is prevented.

The list keeps two weighting measures for each vocabulary. Each one of these two measures represent the closeness/distance of the word to one of the two dialects. At this stage, classifier assigns a value of 100 or 0 as the weighting measure (closeness/distance) to the vocabulary.

During this phase, the training phase, the classifier might find a word that is already in its weighting matrix. If, for example, this word has previously been assigned to Kurmanji, and now it has been found in a Sorani text as well, the weighting entry for Sorani would be set to 100 too. In other words, it means that the word is equally considered as Kurmanji and Sorani. At the end of this phase the required knowledge of the classifier has been generated.

Figure 3 shows a piece of the Weighting List file. In this list, the first column shows the row number. The second column shows the vocabulary. The third column shows the Kurmanji weight of the vocabulary. The fourth column shows the Sorani weight of the vocabulary.

In this sample there is no common words between the two dialects. However, there are commonalities between the vocabularies of these dialects. We have shown this briefly in section 5 (Experiment). The importance of this commonality and how it would affect the NLP in Kurdish is out of the scope of this research.

5302	rêka	100	0
5303	rêkany	100	0
5304	rêknekewtnî	0	100
5305	rêkupêkî	0	100
5306	rêkupêkîyek	0	100
5307	rêkxrawe	0	100
5308	rêkxrawî	0	100
5309	rêkxstneweî	0	100
5310	rêkên	100	0
5311	rêvebirina	100	0
5312	rêwcê	0	100
5313	rêyan	100	0

Figure 3. A sample of Weighting List

It is worth mentioning that the measures of 100 was used for the later developments. At this stage, this seems to be a binary function, returning *true* if a vocabulary belongs to a certain dialect, *false*, otherwise. However, it was not used this way. Instead, it was used as a measure which participated in the dialect classification cumulatively. In fact, we are interested in further research about this case and to assign different closeness/distance weightings that show the affinity of a word to a particular dialect more precisely. Therefore, the Weighting List could be updated in the future to accommodate different values between 0 and 100. This would require more data which must be manually labelled and used in the training of the classifier. Obviously, the other parts of the research environment would remain unchanged.

4.3. Classification Process

In the classification process, first, the classifier reads the input text and extracts what is, usually, called the "features" in the classification context. Again, during this process, the text processor removes all non-letter characters from the text. The feature extraction happens in two steps. First, the text processor tokenizes the text, counts the words, and updates the vocabulary vector by setting the number of occurrences of each word that it finds in the text. Second, it sets a two entry vector by calculating the weight of each entry.

This process has been formulated as below:

$$W^2_{i=1}[i] = \sum_{j=1}^{n} WL[i,j] \times VC(j), \qquad VC(j) > 0 \tag{1}$$

$$DC^2_{i=1}[i] = min\left(\left(W^2_{i=1}[i] \div 100\right), \qquad 100\right) \tag{2}$$

Given:

W is a vector with two entries corresponding to the two dialects. **WL** is the weighting matrix or feature matrix. **VC** is the number of occurrences of each word in the text that has a corresponding entry in **WL**. Each entry of the **DC** vector in (2) shows the percentage (probability of the text being of a specific dialect) that classifier assigns to the text.

The classifier was developed using Ocatve. Octvae was chosen because it was powerful in handling vectors, arrays, and matrices. It was also a proper open source replacement for this kind of experiment, which otherwise should be performed using MATLAB.

5. EXPERIMENT

The Weighting List creation process produced 6792 words. In the testing phase, several pieces of texts were tested for each dialect. Table 1 shows the result.

Table 1. Dialect classification result

Text Dialect	Best Guess		Worst Guess	
Kurmanji	Kurmanji	92%	Kurmanji	52%
	Sorani	26%	Sorani	26%
Sorani	Sorani	91%	Sorani	50%
	Kurmanji	24%	Kurmanji	24%

Table 2 shows the number of words attributed to each dialect alongside the common words among the dialects. It also shows the percentage of the common words to all words and the total words in each dialect.

Table 2. Dialect classification result

Count	Total	Kurmanji	Sorani
Words	6792	2632	4160
Common Words	208	208	208
Percentage	3%	7%	5%

5.1. Analysis and Discussion

The experiment showed that with a reasonable number of vocabulary of about 7,000 entry, the classifier is able to correctly classify the texts that are found in the media. It also showed that in all cases, the classifier assigns a significant magnitude to the dialect that was not the main dialect of the text. This case could have been considered a usual result of the commonality between the dialects with regard to their vocabulary.

To investigate this case, the common vocabularies between the dialects were counted. The common vocabularies were selected based on their weighting measure in the Weighting List. Obviously, the common vocabularies were those with the same weighting measure. Table 2 shows, the percentage of the common vocabularies. It shows that although based on the collected data there is an insignificant difference between the percentages of the presence of the common vocabularies in each dialect, the rate of commonality is far less than the one that is observable in the dialect identification percentage.

This fact is of interest from different point of views. For instance, it shows that Kurmanji and Sorani dialects are sharing common vocabularies that although do not form a large portion of their lexicon, play an important role as the basis for their lexicon structure. In other words, although Kurmanji and Sorani are considered as two dialects that are mutually unintelligible, their case might not be similar to the case of two different languages.

5.2 Issues

There are several issues that our research has not addressed at this stage. We continue our study to investigate these issues that we believe that they are crucial for the advancement of CL and NLP for Kurdish. These issues are listed below:

- What would be the results of the experiment, if we use a "stemmer" during the Weighting List generation and during the classification process?
- How the lack of a standard orthography affects the entire study?
- How the issue of the proper nouns (proper names) [56, 57] and Named Entity Recognition would affect the entire approach?

Regarding the last item, one may suggest that an immediate remedy could be to find those words that start with a capital letter. Unfortunately, this is not an option, because in Kurdish no uniform rule exists about capitalization of the proper nouns. Even if such a rule existed, it could not be applied to Persian/Arabic script.

6. CONCLUSIONS

The researchers of Computational Linguistics and Natural Language Processing for the dominant languages such as English and German, have not focused on the computational dialectology. However, this subject is important in languages with a diverse and linguistically long distant dialects. In some cases, these dialects might be considered mutually unintelligible. For the languages with this specification, automatic dialect classification/identification becomes a necessary part of Language Technology.

Kurdish language includes several dialects. The two widely spoken dialects are Kurmanji and Sorani. These two dialects are considered by professionals and linguists as mutually unintelligible. This research applied an adapted technique of classification based on an adaptation of Support Vector Machine (SVM) approach for dialect classification of Kurdish texts which are written in the mentioned dialects.

The research showed that with a proper vocabulary list that is used to train the system, the text's dialect could be identified with a high degree of accuracy. The experiments also showed that there is a small number of common lexicon that plays an important role in the forming of the context of each dialect.

However, the area of the research has several unexplored topics. For instance, how developing a "stemmer" that is able to find the *stems* of the text words would affect the classification result, both from the efficiency and the accuracy point of views. Esmaili, Salavati, and Datta [30] have

introduced a rule-based "stemmer". We have also developed a "stemmer" with some differences with the one that Esmaili, Salavati, and Datta have developed. However, this "stemmer" has not been incorporated at this stage of the research.

REFERENCES

[1] C. G. Clopper and D. B. Pisoni, (2007) "Free classification of regional dialects of American English," Journal of Phonetics, vol. 35, no. 3, pp. 421–438. [Online].
Available: http://www.sciencedirect.com/science/ article/pii/S0095447006000301.

[2] G. Tuaillon, (1986) "How the French dialectal data enter the Atlas Linguarum Europae," English, Computers and the Humanities, vol. 20, no. 4, pp. 247– 252. [Online].
Available: http://dx.doi.org/10.1007/BF02400111.

[3] S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, M. Saad, and K. Smaili, (2015). "Cross-dialectal arabic processing," in Computational Linguistics and Intelligent Text Processing, Springer, pp. 620–632.

[4] C. Tang and V. J. van Heuven, (2009). "Mutual intelligibility of Chinese dialects experimentally tested," Lingua, vol. 119, p. 24.

[5] P. G. Kreyenbroek and S. Sperl, (1992) The Kurds: a contemporary review. New York: Routledg

[6] Kurdish Academy of Languages, The Kurdish Population, (2008). [Online].
Available: http://www.kurdishacademy.org/?q=node/199 (visited on 10/05/2014).

[7] D. McDowall, (2005). A Modern History of Kurds. New York: I.B.Tauris.

[8] J. Benesty, M. M. Sondhi, and Y. Huang, (2008). Springer Handbook of Speech Processing. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[9] M. Gasser, (2006). How Language Works, Ed3.0. [Online].
Available: http: //www.indiana.edu/~hlw/book.html (visited on 12/26/2015).

[10] The dialects of Kurdish / home, (2015). [Online].
Available: http://kurdish.humanities.manchester.ac.uk/ (visited on 02/20/2015).

[11] J. Huggler, (2001). The world's largest nation without a state seeks a new home in the west, The Independent. [Online]. Available: http://www.independent.co.uk/news/world/europe/the-worlds-largest-nation-without-a-state-seeks-a-new-home-in-the-west-692440.html (visited on 02/20/2015).

[12] A. Burhan, (2011). "Kurds and Kurdish language," Turkish Studies, vol. 6, no. 03, pp. 43–57. [Online]. Available: http://www.turkishstudies.net/Makaleler/1793359710_4_ahmet_buran.pdf (visited on 02/27/2015).

[13] G. Haig and Y. Matras, (2002). "Kurdish linguistics: a brief overview,".

[14] E. R. McCarus, (1960). "Kurdish language studies," The Middle East Journal, pp. 325–335.

[15] F. Hennerbichler et al., (2012). "The origin of Kurds," Advances in Anthropology, vol. 2, no. 02, p. 64.

[16] A. Hassanpour, (1992). Nationalism and language in Kurdistan, 1918-1985. Edwin Mellen Pr.

[17] T. Jügel, (2014). “On the linguistic history of Kurdish,” Kurdish Studies, vol. 2, no. 2, pp. 123–142.

[18] G. Haig and E. Öpengin, (2014). “Introduction to special issue-Kurdish: a critical research overview,” Kurdish Studies, vol. 2, no. 2, pp. 99–122.

[19] D. N. MacKenzie, (1962). Kurdish Dialect: Studies. Oxford University Press, vol. 2.

[20] J. Nebez, (1976). Ziman-i Yekgirtû-i Kurdi (’Towards a Unified Kurdish Language)[in Kurdish]. Bamberg: NUKSE.

[21] M. R. Izady, The Kurds: A concise handbook. Taylor & Francis, 1992.

[22] Kurdish language | Kurdish academy of language. (2014). [Online]. Available: http://www.kurdishacademy.org/?q=node/41 (visited on 02/27/2015).

[23] L. Paul, (2014). KURDISH LANGUAGE, Encyclopaedia Iranica, online edition. [Online]. Available: http://www.iranicaonline.org/articles/kurdish-language-i (visited on 09/20/2014).

[24] M. Leezenberg, (2015). “Gorani influence on central Kurdish,” [Online]. Available: http://www.kurdishacademy.org/?q=node/10 (visited on 05/29/2015).

[25] KAL featured articles. (2014). [Online]. Available: http://www.kurdishacademy.org/?q=ku/book/export/html/5 (visited on 02/28/2015).

[26] C. A. Bedirxan and R. Lescot, (1986). Kurdische Grammatik: Kurmancî-Dialekt. Kurdisches Institut, vol. 1.

[27] Y. Matras and G. Reershemius, (1991). “Standardization beyond the state: the cases of Yiddish, Kurdish and Romani,” UIP-BERICHTE UIE REPORTS DOSSIERS IUE, p. 103.

[28] S. Krauwer, (2003). “The basic language resource kit (BLARK) as the first milestone for the language resources roadmap,” Proceedings of SPECOM 2003, pp. 8–15, 2003.

[29] H. Hassani and R. Kareem, “Kurdish Text to Speech (KTTS),” in Designing for Global Markets 10 Proceedings of the Tenth International Workshop on Internationalisation of Products and Systems IWIPS 2011, 2011, pp. 79–89.

[30] K. S. Esmaili, S. Salavati, and A. Datta, (2014). “Towards Kurdish information retrieval,” ACM Transactions on Asian Language Information Processing (TALIP), vol. 13, no. 2, p. 7.

[31] B. O. Mohammed, (2013). “Handwritten Kurdish character recognition using geometric discretization feature,” IJCSC, vol. 4, pp. 51–55.

[32] The Republic of Iraq - Ministry of Interior - General Directorate for Nationality, (2005). Constitution of Iraq. [Online]. Available: http://perleman.org/files/sitecontents/070708095356.pdf (visited on 02/19/2015).

[33] Kurdistan Parliament voted for draft constitution. (2014). [Online]. Available: http://www.perlemanikurdistan.com/Default.aspx?page=article&id=5593&l=1 (visited on 02/20/2015).

[34] Draft Constitution of Kurdistan Region: Kurdistan Parliament. (2014). [Online]. Available: http://www.perlemanikurdistan.com/files/sitecontents/100809083313.pdf (visited on 02/20/2015).

[35] M. J. Kelly, (2010). "The Kurdish regional constitution within the framework of the Iraqi federal constitution: a struggle for sovereignty, oil, ethnic identity, and the prospects for a reverse supremacy clause," Penn State Law Review, vol. 114, no. 3, pp. 707–808.

[36] K. R. S. Office, (2014). Iraqi Kurdistan Population Forecast for 2009-2020 [in Kurdish], Erbil.

[37] Kurdish, northern, (2015). Ethnologue. [Online].
Available: http://www.ethnologue.com/language/kmr (visited on 02/20/2015).

[38] Kurdish, central, (2015). Ethnologue. [Online]. Available: http://www.ethnologue.com/language/ckb (visited on 02/20/2015).

[39] Kurdish, southern, (2015). Ethnologue. [Online].
Available: http://www.ethnologue.com/language/sdh (visited on 02/20/2015).

[40] The Kurdish language. (2015). [Online]. Available: http://cabinet.gov.krd/p/p.aspx?l=12&p=215 (visited on 02/20/2015).

[41] Kurdistan TV. (2015). [Online]. Available: http://www.kurdistantv.tv/kurs/Home
(visited on 02/20/2015).

[42] Rudaw. (2015). [Online]. Available: http://http://rudaw.net/sorani (visited on 02/20/2015).

[43] KNN. (2015). [Online]. Available: http://knnc.net/ (visited on 02/20/2015).

[44] Kurdistan Parliament [in Kurdish]. (2015). [Online].
Available: http://www.perlemanikurdistan.com/Default.aspx?l=3 (visited on 02/20/2015).

[45] Turkey 'to allow Kurdish lessons', (2014). BBC News. [Online].
Available: http://www.bbc.com/news/world-europe-18410596 (visited on 02/20/2015).

[46] KLPP - main (EN). (2015). [Online].
Available: http://eng.uok.ac.ir/esmaili/research/klpp/en/main.htm (visited on 02/20/2015).

[47] B. Kessler, (1995). "Computational dialectology in Irish Gaelic," in Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics, ser. EACL '95, Dublin, Ireland: Morgan Kaufmann Publishers Inc., pp. 60–66. [Online].
Available: http://dx.doi.org/10.3115/976973.976983 (visited on 02/20/2015).

[48] J. Nerbonne and W. Heeringa, (2001). "Computational comparison and classification of dialects," Dialectologia et Geolinguistica, vol. 9, no. 2001, pp. 69–83.

[49] O. F. Zaidan and C. Callison-Burch, (2014). "Arabic dialect identification," Computational Linguistics, vol. 40, no. 1, pp. 171–202.

[50] K. Nigam, A. Mccallum, S. Thrun, and T. Mitchell, (2000). "Text classification from labeled and unlabeled documents using em," English, Machine Learning, vol. 39, no. 2-3, pp. 103–134, [Online].
Available:http://dx.doi.org/10.1023/A:1007692713085 (visited on 03/22/2015).

[51] J. Burstein, D. Marcu, S. Andreyev, and M. Chodorow, (2001). "Towards automatic classification of discourse elements in essays," in Proceedings of the 39th annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 98–105.

[52] J. Staš, J. Juhár, and D. Hládek, (2014). "Classification of heterogeneous text data for robust domain-specific language modeling," English, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2014, no. 1, 14, 2014. [Online]. Available: http://dx.doi.org/10.1186/1687-4722-2014-14 (visited on 03/22/2015)

[53] A. Danesh, B. Moshiri, and O. Fatemi, (2007). "Improve text classification accuracy based on classifier fusion methods," in Information Fusion, 2007 10th International Conference on, IEEE, 2007, pp. 1–6.

[54] T. Joachims, (1998). Text categorization with support vector machines: Learning with many relevant features. Springer.

[55] S. Tong and D. Koller, (2002). "Support vector machine active learning with applications to text classification," The Journal of Machine Learning Research, vol. 2, pp. 45–66.

[56] Y. Ravin and N. Wacholder, (1997). Extracting names from natural-language text.

[57] G. Walther, B. Sagot, and K. Fort, (2010). "Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish," in International conference on lexis and grammar, [Online]. Available: https://hal.archives-ouvertes.fr/hal-00510999/document (visited on 02/27/2015).

AUTHORS

Hossein Hassani is a lecturer at the University of Kurdistan Hewlêr since 2007. He holds a BSc in Computer (Software), and an MSc in Information Management. He is also a PhD candidate in Computer Science.

Dzejla Medjedovic is an Assistant Professor and Vice Dean of Graduate Program at the Sarajevo School of Science and Technology. She has obtained her PhD in Computer Science from the Stony Brook University.