Subdialectal Differences in Sorani Kurdish

Shervin Malmasi[1,2]
[1] Harvard Medical School, Boston, MA 02115, USA
[2] Macquarie University, Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Abstract

In this study we apply classification methods for detecting subdialectal differences in Sorani Kurdish texts produced in different regions, namely Iran and Iraq. As Sorani is a low-resource language, no corpus including texts from different regions was readily available. To this end, we identified data sources that could be leveraged for this task to create a dataset of 200,000 sentences. Using surface features, we attempted to classify Sorani subdialects, showing that sentences from news sources in Iraq and Iran are distinguishable with 96% accuracy. This is the first preliminary study for a dialect that has not been widely studied in computational linguistics, evidencing the possible existence of distinct subdialects.

1 Introduction

Language Identification (LID) is the task of determining the language of a given text, which may be at the document, sub-document or even sentence level. Recently, attention has turned to discriminating between close languages, such as Malay-Indonesian and Croatian-Serbian (Ljubešić et al., 2007), or even dialects or varieties of one language (British *vs.* American English). LID has several useful applications including lexicography, authorship profiling, machine translation and Information Retrieval. Another example is the application of the output from these LID methods to adapt NLP tools that require annotated data, such as part-of-speech taggers, for resource-poor languages. This will be further discussed in §2.2.

The primary aim of this work is to apply classification methods to regional variants of Central Kurdish, also known as Sorani. Kurdish is a low-resourced but important language. It is classified within a group of "non-Western European languages critical to U.S. national security".[1] In recent years there has been increasing research interest in processing Kurdish (Aliabadi, 2014; Esmaili et al., 2014).

As we will outline in §3, Kurdish is spoken in a number of countries and has several dialects. Sorani is one of these dialects and is spoken in several regions. The main objective of this study is to determine whether subdialectal variations in Sorani can be identified in texts produced from different regions. More specifically, we consider the two main areas where Sorani is spoken, Iran and Iraq.

As the first such study, we identify the relevant data sources and attempt to establish the performance of currently used classification methods on this dialect. We also make available a dataset of 200,000 Sorani sentences to facilitate future research. We approach this task at the sentence-level by developing a corpus of sentences from different regions in §4 and applying classification methods.

2 Related Work and Background

2.1 Language and Variety Identification

Work in language identification (LID) dates back to the seminal work of Beesley (1988), Dunning (1994) and Cavnar and Trenkle (1994). Automatic LID methods have since been widely used in NLP. Although LID can be extremely accurate in distinguishing languages that use distinct character sets (e.g. Chinese or Japanese) or are very dissimilar (e.g. Spanish and Swedish), performance is degraded when it is used for discriminating similar languages or dialects.

[1] https://www.nsep.gov/content/critical-languages

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 89–96, Osaka, Japan, December 12 2016.

This has led to researchers turning their attention to the sub-problem of discriminating between closely-related languages and varieties. This issue has been investigated in the context of confusable languages/dialects, including Malay-Indonesian (Bali, 2006), Croatian-Slovene-Serbian (Ljubešić et al., 2007), Bosnian-Croatian-Serbian (Tiedemann and Ljubešić, 2012), Farsi-Dari (Malmasi and Dras, 2015a) and Chinese varieties (Huang and Lee, 2008).

This issue was also the focus of the recent "Discriminating Similar Language" (DSL) shared task.[2] The shared task used data from 13 different languages and varieties divided into 6 sub-groups and teams needed to build systems for distinguishing these classes. They were provided with a training and development dataset comprised of 20,000 sentences from each language and an unlabelled test set of 1,000 sentences per language was used for evaluation. Most entries used surface features and many applied hierarchical classifiers, taking advantage of the structure provided by the language family memberships of the 13 classes. More details can be found in the shared task report (Zampieri et al., 2014).[3]

Although LID has been investigated using data from many languages, to our knowledge, the present study is the first treatment of Sorani within this context.

2.2 Applications of LID

Further to determining the language of documents, LID has applications in statistical machine translation, lexicography (e.g. inducing dialect-to-dialect lexicons) and authorship profiling in the forensic linguistics domain. In an Information Retrieval context it can help filter documents (e.g. news articles or search results) by language and even dialect; one such example is presented by (Bergsma et al., 2012) where LID is used for creating language-specific Twitter collections.

LID serves as an important preprocessing method for other NLP tasks. This includes selecting appropriate models for machine translation, sentiment analysis or other types of text analysis, *e.g.* Native Language Identification (Malmasi et al., 2013; Malmasi and Dras, 2015b).

LID can also be used in the adaptation of NLP tools, such as part-of-speech taggers for low-resourced languages (Feldman et al., 2006). If Sorani subdialects are too different to directly share the same resources and statistical models, the distinguishing features identified through LID can assist in adapting existing resources for one subdialect to another.

3 Kurdish Language Overview

Spoken by twenty to thirty million Kurds (Haig and Matras, 2002; Esmaili and Salavati, 2013; Salih, 2014; Blau, 2016; Kurdish Academy of Language, 2016), "Kurdish" as a language is nonetheless not easily defined, producing both difficulty and debate for many scholars and researchers (Haig and Matras, 2002; Hassani and Medjedovic, 2016; Paul, 2008). Kurdish, spoken in "Kurdistan" (a region split primarily among Turkey, Iran, Iraq and Syria (Esmaili and Salavati, 2013; Haig and Matras, 2002)), has been embroiled in conflict, so the question of Kurdish linguistic boundaries is complicated by political, cultural and historical factors (Paul, 2008).

One reason for disagreement about the Kurdish language is that Kurdish ethnic identity plays a large role in shaping who claims to speak "Kurdish" (Paul, 2008), and Kurdish ethnic identity is highly politicized (Nerwiy, 2012), especially with regards to a singular "Kurdish language" or plural "Kurdish languages" (Kurdish Academy of Language, 2016; Paul, 2008). While being described as a "dialect-rich language, sometimes referred to as a dialect continuum" (Esmaili and Salavati, 2013), the very act of categorizing a "Kurdish" dialect as a separate language or a separate language as a "Kurdish" dialect is contentious. Further complicating the Kurdish language is its precarious situation in recent history: Kurdish is recognized as an official language in Iraq (Hassani and Medjedovic, 2016), but its pedagogy and use have also been banned in Turkey and Syria (Blau, 2016; Nerwiy, 2012; Salih, 2014).

[2] Held at the Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, co-located w/ COLING 2014.
[3] The task was also expanded and held in 2015 and 2016.

3.1 Geography

Historically, Kurdistan was divided into North and South by the Byzantine and Islamic empires and into Northwest and East by the Ottoman and Persian empires (Nerwiy, 2012). After World War I, Kurdistan was divided among Turkey, Persia, Iraq and Syria (Blau, 2009). Kurds have also lived in Armenia, Lebanon and Egypt for centuries and have formed strong diasporic communities in Europe and North America (Haig and Matras, 2002; Hassani and Medjedovic, 2016).

Most Kurds are bilingual or multilingual, speaking other languages like Arabic, Turkish or Farsi in addition to Kurdish (Salih, 2014). Having been immersed in multilingual contexts for centuries, Kurds—presumably Kurdish speakers—have interacted closely with speakers of Arabic, Armenian, Persian, New Aramaic and Turkish, which have all left traces in the Kurdish language (Haig and Matras, 2002).

3.2 Language Family

Kurdish belongs to the Indo-European family of languages within the northwestern Iranian group (Kurdish Academy of Language, 2016; Nerwiy, 2012; Paul, 2008), though it may also be described as only solidly belonging to a western subgroup of Iranian languages encompassing Persian, Kurdish, Balōči and the other Iranian languages of present-day Iran (Paul, 2008). An important isogloss shared among Persian, Kurdish and Balōči in contrast with many northwestern Iranian languages is the past stem of many verbs formed with the vowel *i* (Paul, 2008).

Still, relations of borrowing and influence within the western Iranian languages—that is, between Kurdish, Persian and Balōči (and Zazaki and Gurani as different from Kurdish, dependent upon the classification system)—are complex and not always straightforward based upon phonological and morphological evidence (Paul, 2008).

3.3 Dialects

Kurdish dialects are highly differentiated and are not mutually intelligible (Paul, 2008). It has even been suggested that the two main dialects (the Northern and Central dialects) of Kurdish can be treated as separate languages (Haig and Matras, 2002).

For both political and linguistic reasons, both Kurds and scholars have disagreed as to how Kurdish dialects should best be taxonomized (Kurdish Academy of Language, 2016; Paul, 2008; Zahedi and Mehrazmay, 2011). The broadest classifications of the Kurdish language include the Northern, Central and Southern dialects as well as Zazaki and Gurani. Other classifications typically disqualify Zazaki and Gurani based upon linguistic evidence and mainly focus upon the Northern and Central dialects (Zahedi and Mehrazmay, 2011), which account for over 75% of native Kurdish speakers (Esmaili and Salavati, 2013). A study put forth by D. N. Mackenzie in 1961—which treats Zazaki and Gurani as separate languages—remains standard in linguistic research and considers "Kurdish" to be a language divisible into the Northern, Central and Southern Kurdish dialects (Haig and Matras, 2002; Paul, 2008).

Of the Northern, Central and Southern dialects, Kurdish dialects differ morphologically in terms of case and gender (Haig and Matras, 2002; Paul, 2008; Zahedi and Mehrazmay, 2011), though researchers have found exceptions to these rules (Esmaili and Salavati, 2013; Haig and Matras, 2002; Hassani and Medjedovic, 2016). Northern Kurdish dialects distinguish gender and retain an inflectional system for masculine nouns (Haig and Matras, 2002; Paul, 2008). Some Central Kurdish dialects also have a case system, but some have also dropped case distinction entirely; all Central Kurdish dialects do not distinguish gender (Haig and Matras, 2002; Paul, 2008). South Kurdish dialects do not distinguish gender, and some show different forms of plural endings (Paul, 2008).

Ezafe in Kurdish dialects are more complex in the Northern dialects, which distinguish a masculine/feminine *-ē* / *-ā* and plural *-ēt*, and simplifies toward the Southern dialects, which have one general form of the Ezafe, *-ī*. Ezafe in the Central dialects can be considered intermediate (Paul, 2008).

Central and Southern Kurdish dialects utilize suffix pronouns, while the Northern dialects do not (Paul, 2008). Central and Southern Kurdish dialects also have a secondary passive conjugation which does not exist in the Northern dialects (Paul, 2008).

3.3.1 Northern Kurdish Dialects

Most Kurds in Turkey, Iraqi Kurdistan, northeastern Syria and the former Soviet Union (especially Georgia and Armenia) speak the Northern Kurdish dialects (also referred to as Kurmanji, Northern Kurmanji or Badinani (Kurdish Academy of Language, 2016; Paul, 2008; Nerwiy, 2012)). The Northern Kurdish dialects encompass around 20 million speakers (Hassani and Medjedovic, 2016) and primarily use a Latin-based alphabet (Esmaili and Salavati, 2013), as we will describe below.

3.3.2 Central Kurdish Dialects

Most Kurds located around Arbil, Suleymaniyeh and Kirkuk in Iraq as well as those in Iranian Kurdistan speak the Central Kurdish dialects (also referred to as Sorani (Kurdish Academy of Language, 2016; Paul, 2008; Nerwiy, 2012)). Speakers of the Central Kurdish dialects number around seven million (Hassani and Medjedovic, 2016) and use an Arabic-based alphabet (Esmaili and Salavati, 2013).

Regarding verb morphology, Central Kurdish dialects employ clitics with verb stems as concord markers, a feature that distinguishes them from the Northern dialects. In the Central dialects, the positions of these markers vary according to negation, auxiliary usage, valency and thematic roles; in the Northern dialects, concord markers are fixed, following the verb stem (Haig and Matras, 2002).

For verbal agreement and alignment, Central Kurdish dialects use pronominal enclitics that attach to the direct object for past-tense transitive constructions, whereas the Northern dialects use ergative constructions (Esmaili and Salavati, 2013; Haig and Matras, 2002).

A further point of distinction is that the Central (and Southern) Kurdish dialects use the definite marker *-aka*, which is absent in the Northern dialects (Esmaili and Salavati, 2013; Zahedi and Mehrazmay, 2011).

This is the dialect being investigated in this study. We're interested in identifying subdialectal differences in Sorani texts sourced from different regions, namely Iran and Iraq.

3.3.3 Southern Kurdish Dialects

The Southern Kurdish dialects (also referred to as Pehlewani, Pahlawanik (Kurdish Academy of Language, 2016; Paul, 2008) or Hawramani (Salih, 2014)) are spoken primarily in the Khanaqin and Mandalin districts of Iraqi Kurdistan and in the Kermanshah region of Iran (Nerwiy, 2012).

3.4 Orthography

Kurdish utilizes four scripts for writing (Latin, Perso-Arabic, Cyrillic and Yekgirtú) dependent upon geographical, political and cultural factors; it lacks a standard, formalized orthographic system (Hassani and Medjedovic, 2016). Nevertheless, there have been efforts at standardization (Haig and Matras, 2002), and most research recognizes the Latin and Arabic scripts as being the most prominent, earning Kurdish the title of being a bi-standard language (Esmaili and Salavati, 2013; Zahedi and Mehrazmay, 2011).

The Central dialects adapted the Perso-Arabic script in the city of Suleymaniya in Iraqi Kurdistan in the nineteenth century, and this script has been used in the Kurdish regions of Iraq and Iran that speak the Central dialects (Haig and Matras, 2002). For the Northern dialects, Kurdish nationalists adapted Arabic script in the nineteenth century; later in the 1930s, Celadet Bedir-Khan introduced a Latin script for Kurdish that has been used in Turkish and Syrian Kurdistan as well as in the European diaspora (Haig and Matras, 2002; Hassani and Medjedovic, 2016). In 1940, the Cyrillic script for Kurdish was converted from a Roman script developed in the Soviet Union and is mainly based upon the Northern dialects (Haig and Matras, 2002). As compared to the Perso-Arabic script and the Cyrillic script, usage of the Latin script is growing (Hassani and Medjedovic, 2016). Additionally, the Kurdish Academy of Language recently proposed Yekgirtú as a unified alphabetic system for Kurdish (Zahedi and Mehrazmay, 2011).

Mapping the Kurdish Latin-based alphabet (used by the Northern dialects) to the Kurdish Arabic-based alphabet (used by the Central dialects) yields twenty-four one-to-one mappings, four one-to-two mappings and five one-to-zero mappings (Esmaili and Salavati, 2013). The one-to-two and one-to-zero mappings attest respectively to the ambiguities of the alphabets and to the differences between the Northern and Central dialects (Esmaili and Salavati, 2013; Hassani and Medjedovic, 2016). Both orthographies are alphabetic, meaning that vowels must be written (Esmaili and Salavati, 2013).[4]

[4]This is in contrast with other abjad writing systems that use the Perso-Arabic script.

4 Data

As Sorani is a low-resourced language, no corpus including texts from different regions was readily available. However, the amount of Sorani language content on the web has been increasing and this provides a good source of data for building corpora.

Similar to the recent work in this area, we approach this task at the sentence-level. Sentence length, measured by the number of tokens, is an important factor to consider when creating the dataset. There may not be enough distinguishing features if a sentence is too short, and conversely, very long texts will likely have more features that facilitate correct classification. This assumption is supported by recent evidence from related work suggesting that shorter sentences are more difficult to classify (Malmasi and Dras,). Bearing this in mind, we limited our dataset to sentences in the range of 5–55 tokens in order to maintain a balance between short and long sentences.

For this study we opted to extract data from news providers based in Iran and Iraq as the source of our data. For Iraq we chose Rudaw and Sharpress, while for Iran we used Sahar TV and Kurdpress. Using articles from these news sources, a total of $100{,}000$ sentences matching our length requirements were extracted for each class, resulting in a corpus of $200{,}000$ sentences. We also make this data freely available to other researchers.[5]

5 Experimental Methodology

We approach this task as a binary classification problem, splitting our data into two classes representing Sorani texts from Iran and Iraq.

5.1 Features

We employ two lexical surface feature types for this task, as described below. The sentences are tokenized based on whitespace and punctuation prior to feature extraction.

Character n-grams This is a sub-word feature that uses the constituent characters that make up the whole text. When used as n-grams, the features are n-character slices of the text. From a linguistic point of view, the substrings captured by this feature, depending on the order, can implicitly capture various sub-lexical features including single letters, phonemes, syllables, morphemes and suffixes. In this study we examine n-grams of order 2–4.

Word n-grams The surface forms of words can be used as a feature for classification. Each unique word may be used as a feature (i.e. unigrams), but the use of bigram distributions is also common. In this scenario, the n-grams are extracted along with their frequency distributions. For this study we evaluate unigram features.

5.2 Named Entity Masking

Our dataset is not controlled for topic and it is possible implicitly capture topical cues and are thus susceptible to *topic bias*. *Topic bias* can occur as a result of the themes or topics of the texts to be classified not being evenly distributed across the classes, leading to correlations between classes and topics (Brooke and Hirst, 2012; Malmasi and Dras, 2014a; Malmasi and Dras, 2015c). More specifically for the current work, the topics can refer to regional toponyms and location names.

One way to counter this issue is to create a balanced or parallel corpus (Malmasi and Dras,). This is a non-trivial task that requires time and resources, and so was not considered for this preliminary study. Another approach is based on named entity masking, which aims to identify and remove named entities such as location names to minimize their influence on the classification models. This approach requires the identification of such tokens through Named Entity Recognition (NER) or some other method. The 2015 DSL Shared Task included evaluation using such masked texts where this was achieved using a heuristic method that masked all capitalized tokens (Zampieri et al., 2015). However, given the lack of NER systems for Sorani and the absence of capitalization info in the Perso-Arabic script, it was not

[5] http://web.science.mq.edu.au/%7Esmalmasi/resources/sorani

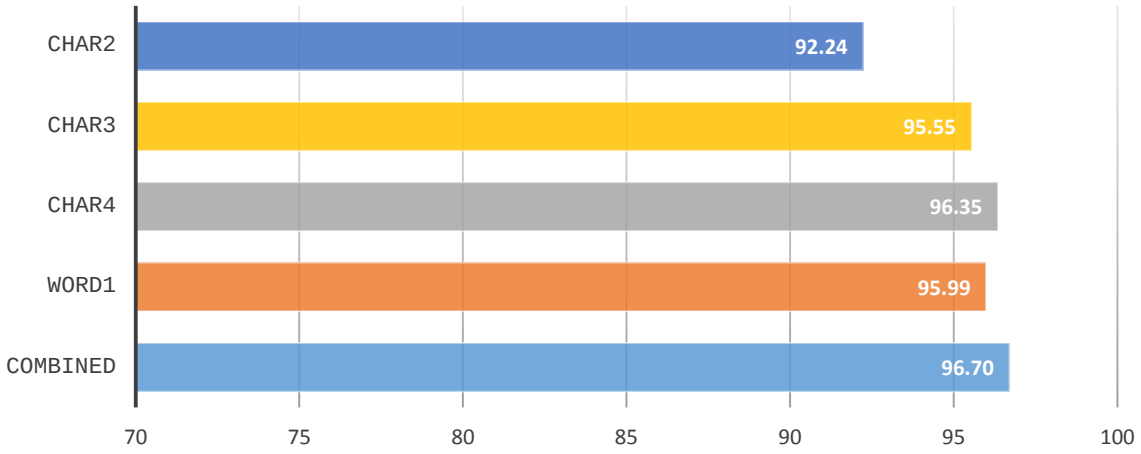

Figure 1: Accuracy for our individual features and their combination.

possible to use either approach. Therefore, in this study we limited our entity masking to the names of the websites and news agencies that we chose as our data sources.[6] More sophisticated entity masking approaches will be considered in future work.

5.3 Classifier

We use a linear Support Vector Machine to perform multi-class classification in our experiments. In particular, we use the LIBLINEAR[7] SVM package (Fan et al., 2008) which has been shown to be efficient for text classification problems with large numbers of features and documents.[8]

5.4 Evaluation

Consistent with most previous studies, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. For creating our folds, we employ stratified cross-validation which aims to ensure that the proportion of classes within each partition is equal (Kohavi, 1995).

We use a *random baseline* for comparison purposes. This is commonly employed in classification tasks where it is calculated by randomly assigning labels to documents. It is a good measure of overall performance in instances where the training data is evenly distributed across the classes, as is the case here. Since our data is equally distributed across both classes, this baseline is 50%.

6 Results

Our experiment explores the classification of Sorani sentences within our corpus using 10-fold cross-validation. We experiment with the different features discussed in the previous section and their combination. The results are shown in Figure 1. All of our features surpass the random baseline of 50% by a large margin. We observe that character n-grams, particularly 4-grams, are very useful here with 96.4% accuracy using a single feature type. Word unigrams are also very informative here with 96.0% accuracy. These results indicate that important lexical and orthographic differences may exist between the Sorani texts from different regions.

We also tested combinations of the features types into a single feature vector, showing that this can yield slightly improved results, with 96.7% accuracy.

It is interesting that character n-grams are a slightly better feature than words. These results also suggest that, at least for this dataset, character n-grams generalize the most. However, it may be the case that word unigrams may perform better with a sufficiently large dataset.

[6] For example: ڕووداو

[7] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

[8] SVM has proven to perform well for large text classification tasks (Malmasi and Dras, 2014c; Malmasi and Dras, 2014b).

7 Discussion and Conclusion

In this study we explored methods for the automatic identification of Sorani subdialects, showing that sentences from news sources in Iraq and Iran are distinguishable with 96% accuracy. This is a new result for dialect that has not previously been experimented with. To this end, we also identified data sources that could be leveraged for this task.

Future work can be directed in several directions. First, detailed analysis of the most discriminative features can provide useful insights about the subdialectal differences. They may reveal interesting sources of influence from Arabic and Persian. A preliminary analysis of the discriminative features showed such differences, but a detailed analysis will be left for future research. While we did not observe a disproportionate amount of named entities in these distinguishing features, methods to eliminate their influence will be important for future work.

Expanding the dataset with additional data from different sources could also be helpful. Further refinement of the dataset to create a topic-balanced corpus can also help conduct more robust experiments in the future. From a machine learning perspective, classifier ensembles have been shown to improve classification performance for numerous NLP tasks. Their application here could also increase system accuracy. Finally, conducting an error analysis on the data could also help better understand the subdialectal differences. Feedback from native speakers would be of assistance here in better documenting the distinguishing features of each dialect, as learned by our models.

Acknowledgements

A special thanks to the reviewers for their helpful comments and feedback.

References

Purya Aliabadi. 2014. Semi-Automatic Development of KurdNet, The Kurdish WordNet. In *ACL (Student Research Workshop)*, pages 94–99.

Ranaivo-Malançon Bali. 2006. Automatic Identification of Close Languages–Case Study: Malay and Indonesian. *ECTI Transaction on Computer and Information Technology*, 2(2):126–133.

Kenneth R Beesley. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, volume 47, page 54. Citeseer.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74. Association for Computational Linguistics.

Joyce Blau. 2009. Kurdish Language ii. History of Kurdish. *Encyclopaedia Iranica*.

Joyce Blau. 2016. The Kurdish Language and Literature. Accessed: 2016-09-20.

Julian Brooke and Graeme Hirst. 2012. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 779–784, Istanbul, Turkey, May.

William B. Cavnar and John M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US.

Ted Dunning. 1994. *Statistical identification of language*. Computing Research Laboratory, New Mexico State University.

Kyumars Sheykh Esmaili and Shahin Salavati. 2013. Sorani Kurdish versus Kurmanji Kurdish: An Empirical Comparison. In *ACL*, pages 300–305.

Kyumars Sheykh Esmaili, Shahin Salavati, and Anwitaman Datta. 2014. Towards Kurdish information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(2):7.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC*, pages 549–554.

Geoffrey Haig and Yaron Matras. 2002. Kurdish linguistics: a brief overview. *STUF-Language Typology and Universals*, 55(1):3–14.

Hossein Hassani and Dzejla Medjedovic. 2016. Automatic Kurdish Dialects Identification. *Computer Science & Information Technology*.

Chu-Ren Huang and Lung-Hao Lee. 2008. Contrastive Approach towards Text Source Classification based on Top-Bag-Word Similarity.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145.

Kurdish Academy of Language. 2016. Kurdish Language. Accessed: 2016-09-20.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language indentification: How to distinguish similar languages? In *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, pages 541–546. IEEE.

Shervin Malmasi and Mark Dras. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *PACLING 2015*.

Shervin Malmasi and Mark Dras. 2014a. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*, pages 180–186, Doha, Qatar. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014b. Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, pages 95–99, Gothenburg, Sweden. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014c. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015a. Automatic Language Identification for Persian and Dari texts. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 59–64, Bali, Indonesia, May.

Shervin Malmasi and Mark Dras. 2015b. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of NAACL-HLT 2015*, Denver, Colorado, June. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015c. Multilingual Native Language Identification. In *Natural Language Engineering*.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June. Association for Computational Linguistics.

Hawar Khalil Taher Nerwiy. 2012. *The Republic of Kurdistan, 1946*. Ph.D. thesis, Faculty of the Humanities, Leiden University.

Ludwig Paul. 2008. Kurdish language. i. History of the Kurdish language. *Encyclopaedia Iranica*.

Rashwan Ramadan Salih. 2014. *A comparative study of English and Kurdish connectives in newspaper opinion articles*. Ph.D. thesis, Department of English, University of Leicester.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634.

Keivan Zahedi and Roghayeh Mehrazmay. 2011. Definiteness in sorani Kurdish and English. *Dialectologia: revista electrònica*, (7):129–157.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. *COLING 2014*, page 58.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the dsl shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, pages 1–9, Hissar, Bulgaria.