Jurnal Informatika
Vol. 17., No. 1, January 2023, pp. 1-17
ISSN 1978-0524 (print) | 2528-6374 (online)
http://journal.uad.ac.id/index.php/JIFO/index

Towards a Complete Kurdish NLP Pipeline: Challenges and Opportunities

Dastan Maulud [a,1], Karwan Jacksi [b,2,*], Ismael Ali [b,3]

[a] Department of Information Technology, Technical College of Informatics-Akre, Duhok Polytechnic University, Duhok, Kurdistan Region – Iraq.
[b] Department of Computer Science, University of Zakho, Duhok, Kurdistan Region – Iraq.
[1] dastan.mawlud@mhe-krg.org; [2] karwan.jacksi@uoz.edu.krd ; [3] Ismael.Ali@uoz.edu.krd
* Corresponding Author

Received 25 October 2022; accepted 20 December 2022; published 10 January 2023

ABSTRACT

With the rapid growth of Kurdish language content on the web, there is a high demand for making this information readable and processable by machines. In order to accomplish this, the Kurdish Natural Language Processing (KNLP) pipeline is required. Computers that can process human language use the field of Natural Language Processing (NLP). In its efforts to bridge the communication gap between humans and computers, NLP draws from a wide range of fields, including computer science and computational linguistics. There have been some notable efforts made toward creating the KNLP pipeline. However, it does not support the complete NLP tasks needed to enable semantic web and text mining applications. This paper surveys the work done in the field of NLP for the Kurdish language, its applications, and linguistic challenges.

KEYWORDS
Text Corpus
Annotated Corpus
Kurdish Language
NLP
Semantic Web
Text Mining

1 Introduction

The amount of Kurdish text is increasing online, and as any other language, it is gaining a high demand for making this web content readable and processable by machines. However, in terms of computing there is a need for the Kurdish language to be understood and produced by technology as it is the case for other non-Kurdish text on the web [1]. The future of NLP in Kurdish language development may be explored by deepening the research on a diverse range of tasks and advances in the NLP pipeline [2], [3].

Kurdish language is a member of the Indo-Iranian branch of Indo-European languages which is spoken by more than 30 million people in Western Asia, mainly in Iraq, Turkey, Iran, and Syria [4], [5]. The Kurdish language has a variety of dialects and owns its own grammatical system and rich lexicon [6]. Kurdish has traditionally been written in a variety of scripts, as a result the Kurmanji dialect is predominantly written in Latin, whilst Sorani, Southern Kurdish, and Laki are predominantly written in modified Arabic alphabet [4], [7]. This does not only complicate communication between readers and speakers, but also adds the difficulty to the language computing [8], [9]. The Fig. 1 illustrates the Latin- and Arabic-based Kurdish alphabets with the International Phonetic Alphabet (IPA) that are utilized in all dialects. [1].

ISSN 2086-8138 (print) | 2745-7249 (online)
Jurnal Informatika
Vol. 17., No. 1, January 2023, pp. 1-17

IPA	b	t͡ʃ	d͡ʒ	d	f	g	h	ʒ	k	l	ɬ	m	n	p	q	f	r	s	ʃ	t	v	w	x	j	z	ʕ	ħ	ɣ	ʔ
Latin	b	Ç	c	d	f	g	h	j	k	l	ł/ll	m	n	p	q	r	ř/rr	s	ş	t	v	w	x	y	z	ʻ/ʼe/ë	ḥ/ʻh	x̄/x	ʼ
Arabic	ب	چ	ج	د	ف	گ	ه	ژ	ک	ل	ڵ	م	ن	پ	ق	ر	ڕ	س	ش	ت	ڤ	و	خ	ی	ز	ع	ح	غ	ئ

(a) Consonants

IPA	aː	æ	eː	ɪ	iː	oː	uː	ʊ	ʉː
Latin	a	e	Ê	i	î	o	û	u	ü
Arabic	ا	ه	ێ		ی	ۆ	وو	و	ۉ

(b) Vowels

Fig. 1. A comparison of the Kurdish alphabets

Kurdish is a strongly derivational language, due in part to its abundance of affixes and problems or challenges [10]. While Sorani lacks gender and grammatical cases, it does contain a complete article marking system for definite, indefinite, and demonstrative nouns in singular and plural forms [11]. Generally, Kurdish has a *subject-object-verb* word order and is a no-subject (or pro-drop) language in terms of grammar. Within dialects and subdialects, the presence of grammatical markers for nominative and oblique cases varies [12]. Another distinguishing characteristic of the Kurdish language is the morphosyntactic alignment of past-tense transitive verbs. In these tenses, the intransitive verb's subject behaves like the transitive verb's past patient [13]-[15].

On the other hand the natural language content on the web needs to be readable and process-able by machines [2], [16] by the NLP pipeline to help grasp, perceive, and control human language [3], [17]. NLP involves various tasks, ranging from low-level tasks such as sentence segmentation, to high-level tasks such as semantic annotation and opinion mining. The Semantic Web is about applying semantics, i.e., context, to data on the Web, to make the web pages easier to process and manage by machines [2], [18]. This paper surveys the work done in the field of NLP for the Kurdish language. The rest of this paper is organized as follows. In Section 1, we first briefly introduce the Kurdish language and its two main dialects then underline their differences from a rule-based perspective. Our methods and procedures for summarizing the articles are discussed in Section 2. Section 3 describes a KLPT literature review and experimental comparison. Section 4, provides the KNLP Applications Literature Review, Kurdish language challenges are discussed in Section 5, the paper's discussion in Section 6, finally, we conclude the paper in Section 7.

2 Method

While Kurdish is a language with few online resources and is at the beginning level in the area of NLP, we have attempted to summarize almost all studies that have worked on Kurdish NLP. There are several dialects of Kurdish, with Sorani and Kurmanji being the most frequently spoken. Thus, all study has been conducted on these two dialects. To have a better understanding of the present state of Kurdish NLP and computational linguistics, for this reason, the reviewed articles have been split into two main sections with comprehensive analysis of the reviewed articles' quality and the limitations of the study.

First section present the state of the art for KNLP accordingly for every step in the standard NLP pipeline. According to this section of the literature review, researchers have used a wide range of methodologies and strategies to address a variety NLP steps, from 2010 to 2020. Table 1 compares the topics studied works in the literature. As given in Table, there are different NLP tasks have been handled and methods been utilized such as text cleaning, tokenization, POS tagging, lemmatization/stemming, and named entity recognition.

In second section we reviewed the publications that provide the software application of the work done in the KNLP. This section demonstrates different types of algorithms, methods and applications in designing the system for implementing and use of different tasks of KNLP. As was shown in the literature review, variety research methods and approaches have been employed to solve a wide range of issues. The following research used Kurdish NLP applications, ranging from (2011) to the present (2021). A summary of the

latest applications of Kurdish language NLP studies is shown in Table 2, As outlined in Table, a series of activities may be addressed using a variety of techniques, including machine learning models, N-gram models, rule-based machine translation, and classification approaches for both Kurdish dialect Sorani and Kurmanji, for the purpose of evaluating the approach performance, the usage of multiple datasets based on the aim is required. Overall, the author concentrated on three distinct fields: dialectology, speech recognition, and machine translation.

The development of KNLP appears promising, and interesting results have been presented for a limited number of NLP steps, but it is not without challenges. As a result, we attempted to identify and classify KNLP's notable limitations and difficulties.

3 KNLP: Literature Review

The Kurdish NLP pipeline is a collection of tools and algorithms created expressly to process and examine text written in the Kurdish language. Various natural language processing operations are included in the pipeline that are important for allowing semantic web and text mining applications. Some of the NLP tasks that the Kurdish NLP pipeline supports: Tokenization, Part-of-speech tagging, Named entity recognition, Dependency parsing, and Sentiment analysis. Applications for the semantic web and text mining can all benefit from these NLP tasks. For instance, structured data from unstructured text can be extracted using named entity recognition and dependency parsing, and this data can subsequently be utilized to fill databases or knowledge graphs. Customer feedback or social media data can be analyzed for patterns using sentiment analysis. In conclusion, the Kurdish NLP pipeline offers a number of NLP operations necessary to enable semantic web and text mining applications. These include dependency parsing, named entity recognition, part-of-speech tagging, sentiment analysis, tokenization including machine translation and speech recognition.

In this section we present the state of the art for KNLP accordingly for every step in the standard NLP pipeline as presented in the Fig. 2 :

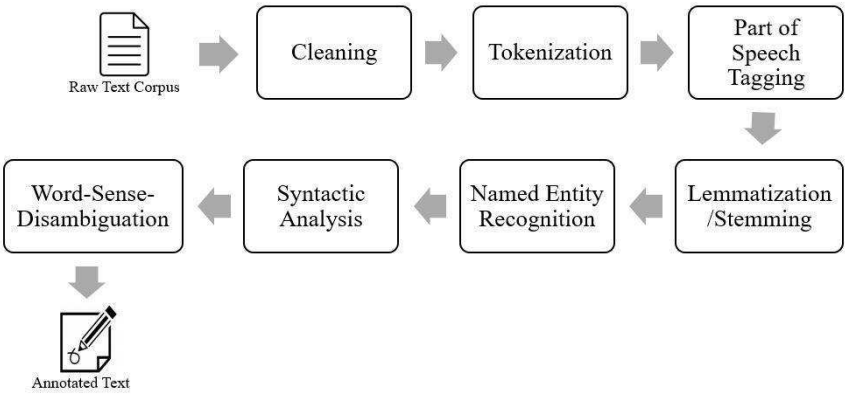

Fig. 2. NLP Pipeline [32].

3.1 Text Cleaning

The very early step in NLP and text mining is to exclude any non-semantic parts of the text, such as removing punctuation marks and stop-words.

S. Ahmadi in [1] the two-function preprocessing module that was given to the Sorani and Kurmanji dialects of the Kurdish language normalized encoding tasks by unifying characters so that each grapheme uses only one unique encoding, and also normalized the orthography of the text. The method unify-numeral is also given to help convert numbers, particularly in Farsi (۰۱۲۳۴۵۶۷۸۹), Eastern Arabic (٠١٢٣٤٥٦٧٨٩) and Western Arabic (0123456789). The author developed a basic system for all scripts; however, users will have the option to customize the numbers in the Kurdish regions. Within the preprocess method, all three of these functions are called upon, with the text normalized, standardized, and unified according to the provided inputs.

A. M. Mustafa et.al. in [19] given a list of stop words for Kurdish Sorani words that are removed from a manuscript once the stemming process begins. A preset list can be created to include these terms that are not necessary for information retrieval but are often used in Kurdish writings. A table lists about 240 stop terms, and for two major reasons the list of stop words is developed: (i) Words that match the phrase and the document must be maintained. It depends heavily on the words which have extreme significance. Noise words should thus be deleted. Documents using phrases such as "بۆ","bo" meaning to "why", "ئێوە", "êwe" meaning "you" and "ئێو", "nêw" meaning "in" the same request must not provide a relevant understanding. These sound words are irrelevant and can harm the efficiency of the retrieval, as they do not differentiate between the relevant documents and those which are not relevant. (ii) In addition, in Kurdish Sorani, the richness of stop words increased the characteristic vector size. Two sources have collected the data: A compilation of information from both television sites, containing 1960 pieces of text information and a total of 43594 words, was compiled by Rudaw, NRT websites. The provided technique is aimed at lowering file size by 35% to 50% by eliminating stop words. In the best of our understanding and study we found that the stop words list was not adequate and that they were not large enough to contain all the stop words in Kurdish.

3.2 Tokenization and Sentence Splitting

The input raw text can be segmented into words or sentences to make it ready for further analysis and processing. The main two segmentation techniques are word-based tokenization and text splitting for the sentence detection task for sentence-based tokenization.

S. Ahmadi in [20] a lexicon and a morphological analyzer are used to tokenize the Sorani and Kurmanji dialects of Kurdish. The authors illustrate how tokenization may effectively solve the language's morphological complexity and absence of a consistent spelling. They use the *WordPunct* tokenizer of NLTK to construct a baseline model. The suggested four-step approach, which includes text preprocessing, compound word tokenization, word tokenization, and morphological analysis, tokenized the text into a series of alphabetic and non-alphabetic letters using a regular expression. The authors annotate 100 sentences from the Kurdish Textbooks Corpus (KTC) [21] for Sorani and 100 sentences from the Pewan corpus [22] to produce a gold-standard dataset to evaluate the effectiveness of the proposed tokenization method. The outcome indicates that tokenization of individual words had a 30.44% accuracy, but the accuracy of compound phrases was either 100% or 0% depending on whether a whitespace was inserted between component parts 100% precision is achieved without white space. The drawback of the study is that compound verb tokenization, along with tense, aspect, person, and mood, verbs are inflected according to the patient, or object of transitive verbs, and may include additional suffixes such as وە(-ewe) to indicate repetition and یش/ش (=îş/=ş) to indicate emphasis.

3.3 Part Of Speech Tagging

Part-of-Speech (POS) tagging is the process of associating words with their respective parts of speech, for example, noun, verb, and adjective. POS tagging is a pre-requisite for more advances NLP tasks such as syntactic analysis.

G. Walther et.al. in [23] based on three sources of data: lexical information, a non-formalized reference grammar, and raw corpora. The authors have created a morphological lexicon and a POS tagger for Kurmanji Kurdish (KurLex), as the following actions were involved: (i) A list of Kurmanji categories is compiled using data from the reference grammar. Simultaneously, they formalized Kurmanji morphology using the lexical formalism, Alexina. (ii) The Kurmanji Kurdish Morphological Lexicon (KurLex) was created by extracting lemmas from several lexical information sources and inflected them. (iii) A 36-tag POS tagset is designed after narrowing the categories list. (iv) Using a variety of simple statistical techniques and heuristics, authors developed different models, then used automatic POS annotation to create a POS-annotated corpus using just lexical information. (v) Finally, this corpus and KurLex were used to train the Maximum-Entropy Lexicon-enriched Tagger (Melt) tagger [24]. PrefHeuristics is the most accurate model, with an

accuracy of 87.5 percent. It's worth mentioning that the limitation of this study is that it is an insufficient research corpus from which to make reliable conclusions.

S. Ahmadi et.al. in [10] analyzed the morphology of the Kurdish language (Sorani dialect) using computational methods to create finite state transducers (FSTs) that were used to generate and interpret words. The described method has four categories, including verbs, nouns, adjectives, and adverbs, of Kurdish morphology. The drawback of the study is the absence of syntactical characteristics that may be used to change morphological forms inside a phrase.

3.4 Lemmatization and Stemming

Text normalization techniques like stemming and lemmatization turn different forms of a word into a single token. This makes the text ready for further word-based processing and analysis.

M. Saeed et.al. in [25] the "Reber" technique is suggested for Sorani dialects of Kurdish to reduce lengthy affixes and prefixes regardless of their sequence. The approach covers three steps: Step one employs a "for loop" that iterates three times, with each loop removing one prefix. After that, three array lists are employed for this purpose, with each resulting string then being compared within the array list to identify the smallest one. The second phase involves removing suffixes using one "for loop" made up of four iterations, with each iteration removing one suffix. This method was developed using Java programming and was used for the KDC-4007 dataset (a dataset with eight classes). For classification, Support Vector Machine (SVM) and Decision Tree (DT) are used. A comparison has been made between the suggested stemmer and the Longest-Match stemmer approach. The F-measure of the Reber stemmer and the longest-match technique in SVM are greater than DT. SVM stemmer with Reber were the best in F-measure, but all other stemmers were lower in Longest-Match. DT stemmers for classes (religion, athletics, and art) using Reber exhibited a greater F-measure, whereas for the remaining classes, F-measure was lower. Even if the authors have gotten promising results, it will be worth trying to replace standard classification methods like neural networks with more advanced ones like deep neural networks on the same dataset, unfortunately, the F score is not written.

S. Salavati et.al. in [26] Jedar, a rule-based stemmer for Sorani Kurdish and Kurmanji Kurdish, was introduced. The authors also employed a state-of-the-art statistical stemming approach, Graph-based RAS (GRAS), and applied it to both of the Kurdish languages. Complete experimental research was then undertaken to assess the efficiency of these stemmers. Jedar manages nested prefixes by using a recursive technique. In addition, the authors have developed two strategies to address Jerad's over-stemming problem. The first strategy used from [27] is to avoid excessive stemming by specifying a minimum stem length parameter. The second strategy is to make use of the Kurdish language's natural suffixing characteristics. Research results [26] suggest that stemming can boost the retrieval of Kurdish papers by up to 35%. They also say that the benefits of the rule-based and statistical methods are equal. In the study, there are no suggestions for how to fix some systemic stemming mistakes, such as over-stemming and bad handling of named entities.

S. Salavati et.al. in [28] proposed a fundamental language processing tool for the Sorani Kurdish language based on the morphological rules and an n-gram language model called Peyv, to extract lemmas of words. These rules have been extracted for different parts of speech, such as nouns, verbs, and adjectives. The authors implemented the Peyv lemmatizer for nouns and verbs, respectively. In the case of nouns, a pruning method is used to find the root. On the other hand, a bottom-up method is used for verbs that have more complex structures than nouns. To evaluate the performance of the proposed tool, the authors analyzed 18M words from 115K news articles from the Pewan text corpus. The accuracy of the Peyv lemmatizer was 86.7%. One of the conclusions that can be drawn from this study is that writing mistakes have previously caused researchers to discover incorrect roots for certain words. Therefore, for developing a more accurate lemmatizer, large number of Kurdish lemmas should be utilized.

A. M. Mustafa et.al. in [19] a Kurdish stemming-step was developed, which is employed for eliminating affixes in the Sorani dialect of the Kurdish language. This is a step-based

method to explain how words go through stages before arriving at the extracted root. The Kurdish stemming-step module aims to catch possible roots by stripping prefixes, suffixes, and postfixes from the input word. The provided word will be verified through all of the Kurdish stemming processes in order to map the string of letters at the beginning or end of the word's root. This technique relied on the creation of sets of possible prefixes and suffixes, which are often used in Kurdish text documents. The proposed approach does not rely on a dictionary for root word checking. For example, (لەهەنگاوەکانیان) becomes ('هەنگاو' means 'phase') in step 1, which drops prefixes (لە) and then does further mapping through several steps, ending with (هەنگاوەکان). The strategy uses the suffixes in the step it follows to identify a match, then it eliminates that suffix from the term at step 10. The prefix (کان) is discarded as a result. After that, the final stage sees the final suffix (ە) matched and deleted. This method is also employed in languages other than English, and not only is it used to remove affixes from nouns and verbs, but it also removes affixes from common stop words used in Sorani Kurdish. A Kurd-specific stemming module was used for the papers gathered, which include a 1960-word text database. Kurdish stemming-step module F measures are close to 1.0, with an average recall of 93 percent. The study's primary limitation is that it cannot manage modest under-and over-stemming mistakes and the precision figure should be given too. It doubt high F measure if recall is 93 percent.

S. ahmadi, in [1] developed morphological rules using Kurdish morphemes in the Sorani and Kurmanji dialects, as well as an annotated lexicon with stem and part-of-speech tags. The proposed module has two classes: stem and spellcheck. While these two groups of classes are each dedicated to separate objectives, they are nevertheless supplied by the same module. The Stem class has four primary functions: stem for extracting words from their stems; lemmatize for lemmatization; analyze for morphological analysis; and suffix_suggest for returning all potential suffixes for a given lexeme. The proposed approach is implemented in Python. The proposed approach is rule-based, which is in turn based on the used dictionaries and corpus, whereas a more generic statistical model is going to be more efficient, at least theoretically.

3.5 Named Entity Recognition

Named entities are definite noun phrases that talk about specific people, places, or ideas, like organizations, people, and dates. Extracting named entities can help a wide range of applications in text mining, information retrieval and ontology engineering.

H. Hassani in [29] a technique has been developed for identifying proper nouns in Sorani and Kurmanji Kurdish texts, by use of rule-based approaches applied to a pair of name dictionaries, a gazetteer, a set of trigrams[1] taken from an untagged corpus, and a limited set of hand-crafted rules. The author proposed a tripartite architecture. (i) Name lists consist of a gazetteer (which includes a subset of proper names as in most existing NER methods), a dictionary of Kurdish names (which includes human names with mostly Kurdish origins), and an Arabic name list (which includes Arabic names that may be recognized in Kurdish texts). (ii) A collection of hand-crafted criteria that include phrases that may occur before or after a proper noun, increasing the likelihood of the candidate names being taken as proper nouns. For instance, if a candidate name is followed by any member of the subset ("aga", "axa", "beg", "xatûn", "xanim"), then the candidate's name is deemed to be a discovered name. (iii) A collection of methods for the recognition of Kurdish names. The proposed approach was tested on 15 documents of various sizes; 8 of them were in Kurmanji and 7 were in Sorani. The precision of the approach they applied was better than 95 percent, and the recall ranged from 40 percent to 80 percent, with the F-measure falling between 60 percent and 80 percent. Due to a lack of name lists, their recall precision was low. The study's limitations include that the name lists were not comprehensive enough to cover the great

[1] A POS string that is also a non-proper-name character. Kurdish Latin script can signify "جمال" (cemal) and can be translated as "beauty" or "face," depending on the context. The suggested method included such names in the Kurdish names dictionary, allowing the algorithm to apply rules depending on the surrounding terms (trigrams).

majority of Kurdish names, and that geographical and place names were not included in the evaluation procedure.

P. Littell et.al. in [30] as part of a pilot research on linguistic rapid response to emergency humanitarian assistance circumstances, it discusses the development of a named-entity recognition (NER) system for the Kurdish and Tajik languages, a dialect of Kurdish. People, places, and organizations were identified in the text. The framework is a conditional random field (CRF) based system that uses L1-regularization to disambiguate ambiguous Sorani forms (also known as Lasso regression). The performance of the proposed technique was evaluated by utilizing the annotated NER data included in the less commonly taught languages (LCTL) language pack [31]. Their results reveal that when adding features, the method's precision is above 74%, recall is 41%-51%, and F-measure is near 51% to 60%. For inferred morphological analyses, authors used input from human linguists, and as we can see, systems perform badly, while adding features the experimental results are low.

3.6 Syntactic Analysis (Relation Recognition)

Syntactic analysis analyzes the sentence grammatically. Modern parsing methodology called dependency parsing (DP) is commonly utilized. The basic principle of DP is that each word is related to every other word via a directed connection. In linguistics, these linkages are termed dependencies. There are a notable efforts performed in the contemporary parsing community [32]. To the best of our knowledge and investigated literature, there has been no work done on syntactic parsing of Kurdish text. The reasons for this can possibly be referred to the challenges mentioned in the section (4).

3.7 Word-Sense-Disambiguation

Word sense disambiguation (WSD) is the task of finding out which "sense" (meaning) of a word is triggered by its use in a certain context, and it appears to happen without people being aware of it [33], such as the work *bank* (I deposited my money in the *bank* close to the *bank* of the city river) which means the financial institution in the first mention and river-side in the second one. The WSD is an issue of assigning words to their correct meanings. In this case, each word is defined by a dictionary and is assigned to its most accurate sense. The context (such as surrounding words) can be used to classify words [34].

Table 1. An overview of the most recent Kurdish language NLP papers.

Ref	NLP Task	Approach	Kurdish Dialect	Application (Field)	Dataset	Results and Accuracy	Limitations
[1], 2020	Text Cleaning	Orthographic conventions	Sorani and Kurmanji	Information retrieval, transliterate	Pewan corpus	Not Mentioned	Rule-based approach reduces accuracy.
[19], 2016	Text Cleaning	Predefined list of stop words	Sorani	Information retrieval, text mining	Rudaw, NRT websites	58%	Lack of Kurdish stop-words.
[20], 2020	Tokenization	Lexicon, morphological analyzer	Sorani and Kurmanji	Information retrieval, text mining	KTC, Pewan corpus	30.44%	Inaccurate single-word tokenization
[10], 2020	POS Tagging	Finite state transducers	Sorani	Morphologica, syntactic analysis	Wergor corpus	Not Mentioned	Syntactical absences can modify phrase morphology.
[23], 2010	POS Tagging	POS tagging categories	Kurmanji	Morphologica, syntactic analysis	Raw corpora	F-Precision 85.7%.	Small sized corpus.
[1], 2020	Lemmatization and stemming	Step-based approach	Sorani	Text mining, transliterate	1960 pieces of text data	Not Mentioned	Statistical inference model to replace the generic rule-based method.
[25], 2018	Lemmatization and stemming	Loops, SVM and DT	Sorani	Information retrieval, text classification	KDC-4007	Not Mentioned	Worthy to attempt deep neural classification algorithms.
[28], 2018	Lemmatization and stemming	Rule-based, the statistical approaches	Sorani and Kurmanji	Morphologica, syntactic analysis	Pewan corpus	86.7%	Small sized corpus

[19], 2018	Lemmatization and stemming	Morphological rules	Sorani	Information retrieval, text mining	Pewan corpus	Precision 86%0 Recall 93%, F-measure 89%	Unable to manage under-and over-stemming mistakes.
[26], 2013	Lemmatization and stemming	Rule-based stemmer, GRAS	Sorani and Kurmanji	Information retrieval, text mining	Pewan corpora	Not Mentioned	Over-stemming and not clear handling of named entities.
[29], 2017	Named Entity Recognition	Rule based, hand-crafted rules	Sorani and Kurmanji	Information retrieval, text mining	online documents	Precision 95% Recall 60% F-measure 80%.	Not examining the most of Kurdish names and geographical and place names.
[30], 2016	Named Entity Recognition	CRF, Lasso regression	Sorani	Information retrieval, text mining	LCTL language pack	Precision 75% Recall 69%-74%, F-measure 51%-60%.	Features improve human linguistics-based inferred morphological analysis.

According to the literature review above, researchers have used a wide range of methodologies and strategies to address a variety NLP steps, from 2010 to 2020. Table 1 compares the topics studied works in the literature. As given in Table, there are different NLP tasks have been handled and methods been utilized such as text cleaning, tokenization, POS tagging, lemmatization/stemming, and named entity recognition.

All papers focused on two most widely spoken dialects, Sorani and Kurmanji. Papers [1], [19] used various methods for performing text cleaning and evaluated the effectiveness of those approaches by applying them on diverse datasets; the output of the approached module in paper [1] may be entered by the user into other modules as an input.

The author in [20] have used Lexicon and a morphological analyzer to tokenize words and for evaluating the approach the Kurdish textbook corpora (KTC) and Pewan corpora are used. It is worth mentioning that, the compound words are not addressed efficiently; whereas the accuracy for compound words is either 100% or 0% depending on whether a whitespace was inserted between component parts.

The authors in [10], [23] have used finite state transducers and a list of POS tagging categories for implementing POS taggers in the field of morphological and syntactic analysis using different datasets for evaluating approaches, and both the suggested methods will enable the assessment of future machine transliteration and machine translation systems.

The authors in [1], [25], [26], [28] have evaluated several techniques for lemmatization and stemming on diverse datasets utilizing many methodologies. One of the primary distinctions between Kurdish and other languages is the fact that Kurdish has several prefixes and suffixes following each other. Some techniques are helpful for various kinds of tasks, including spell checking in text editors. Based on our investigation of the literature, we determined that Kurdish morphology stemming algorithms that are based on the most frequent affixes in the language will be more successful than those based on the most frequent n-grams in each set.

The authors in [29], [30] have worked on named entity recognition tasks for information retrieval and text mining using a variety of methods, including rule-based methods and conditional random fields, while the authors of the paper [29] concentrated on noun identification alone, while the authors of the paper [30] concentrated on individuals' names, locations, and organizations in textual data. The experiment shown that the result obtained in paper [29] is superior than that obtained in paper [30] because it employs rule-based techniques and concentrates on a single NER category.

4 Applications of KNLP: Literature Review

To have a better understanding of the present state of Kurdish NLP and computational linguistics, we reviewed the publications that provide the software application of the work

done in the KNLP. This section demonstrates different types of algorithms, methods and applications in designing the system for implementing and use of different tasks of KNLP.

R. Azad et.al. in [35] developed Kurdish fake news corpus which contains two different sets (crawled fake news from illegitimate sources and manipulated text). The authors applied five machine learning models on the corpus after using Term Frequency-Inverse Document Frequency (TD-IDF) vectorizer as feature selection. The outcomes indicated that the accuracy of the Support Vector Machine (SVM) for set 1 was 88.71% and the Random Forest scored 79.08% for set 2. Although this work showed promising results but due to the limitations in a low-resourced language such as Kurdish there are still many unanswered questions and gaps in other categories.

A. K. Al-Talabani et.al. in [36] developed a technique for identifying dialects and languages by using phonetic and stylistic characteristics. Authors recommended a one-dimensional local binary pattern (LBP) for use in the features: Their research shown that the suggested LBP feature set of the dialect and language recognition systems is effective. The main purpose of this study is to determine how closely related each Kurdish dialect is to its neighbors. Three Kurdish dialects (Sorani, Kurmanji, and Hawrami) and three neighboring languages were used in this study (Arabic, Persian, and Turkish). The conclusion is that the Sorani and Kurmanji dialects are more closely related to each other than the Hawrami dialect is to either, and the closest language to them is Arabic. While Persian is the closest language to the Hawrami dialect. The study's primary limitation is that was no statement of the degree of similarity between dialects and another limitation it could be to the type of similarity used for comparison. Since Kurdish and Arabic belong to very different language family.

H. Hassani et.al. In [37] used a supervised machine learning SVM classification approach to detect the dialects of the Kurdish texts. The research has been done on the Kurmanji and Sorani dialects of Kurdish. The modified approach was used in multiple steps: data collecting, transliteration, and creation of weighting lists. The research showed that when a suitable vocabulary list is utilized to train the system, the dialect of the text can be accurately detected. The research also concluded that there is a limited vocabulary that serves an important function in the creation of each dialect's context. There are some limitations to the research that were not addressed, including the use of a stemmer while generating the weighted list and during classification.

K. M. Kaka-Khan in [38] presented a rule-based machine translation system developed to translate simple English sentences to Kurdish based on the Apertuim free open-source engine [39], which is a rule-based translation tool. The proposed system is considered a good step toward more advanced machine translation tools and technologies; however, the proposed work can only be used to translate some simple sentences, compound sentences, phrases, and idioms from English to Kurdish.

F. Mohammed et.al. In [40] developed a model based on n-grams for categorizing Kurdish Sorani text with many n-gram levels. The dice method was used to categorize the texts, and performance was measured using recall and accuracy. Their results indicated that level 5 n-grams surpass the other n-gram levels in categorization. Despite the promising results of this study, the authors' evaluation of the text classification system is limited to a small number of documents, since they have not stemmed the terms in this work.

H. Hassani in [41] Intralingual machine translation was used to translate documents from Kurmanji to Sorani using word-for-word translation (literal or direct translation) across the dialects. The author employed a modified technique in which they developed a word collocation list as proposed by Zhang [42]. The translated texts were found to be comprehensible in 71% and 79% of cases, for Kurmanji and Sorani, respectively. They are understood in 29 percent of Kurmanji instances and 21 percent of Sorani cases. In order to provide a better translation between the two dialects, the author employed bi-dialectal dictionaries instead of a parallel corpus.

S. Malmasi in [43] Classification methods employed in Kurdish texts written in different areas, notably Iran and Iraq, for finding sub-dialect variations in Sorani. Utilizing surface characteristics n-grams of the character and n-grams of the word. Their findings indicated that sentences from Iraqi and Iranian news sources may be distinguished with 96% accuracy.

This work resulted in novel results however it was limited to the Sorani dialect rather than covering other Kurdish dialects as Kurmanji.

H. Hassani et.al. in [44] developed a Kurdish Text-to-Speech software based on Concatenative Synthesis method. The Concatenative method uses diphone units to have a better transition between the phonemes and to have a more natural speech. The system has been tested by a group of volunteer listeners, and the result of the test showed that the produced speech has a good score of intelligibility, however the work on the reverse process of Kurdish Speech-to-Text has to be addressed by the KNLP research community due to its wide range of applications in high-tech applications as in Internet of Things.

H. K. Hamarashid et.al. in [45] proposed a new word recommendation method for Sorani and Kurmanji Kurdish languages. The Stupid-Back Off method was implemented in the suggested system, and the N-gram model was employed. The proposed system includes the following steps: data collecting, reading the Kurdish text corpus in R-studio, encoding the text corpus, data cleaning, producing N-grams, saving the N-grams, changing characters or letters, reading the stored N-grams, and predicting the next words. The accuracy of the proposed model was high as 96.3% on the used datasets. This work can be more valuable if testing it on a variety of texts in the terms of domains of knowledge such as texts from sport, economics, technology and medicine.

A. Qader et.al. in [46] the Basic Dataset for Sorani Kurdish Automatic Speech Recognition (BD-4SK-ASR) is used to create an automated speech recognition system for Sorani Kurdish. The project's aim was to build a system capable of automatically recognizing basic phrases based on the language used in grades one through three of elementary schools in Iraq's Kurdistan Region. The authors utilized CMUSphinx to train the system and created a (BD-4SK-ASR) dataset. In addition to the study's shortcomings, a lot more work is required to create Kurdish ASR. There are many areas of study that require investigation, and a lot of resources must be acquired and utilized. One difficulty in expanding the use of the Kurdish ASR, or language resources, is a need for a bigger corpus of data, along with study on the languages of other Kurdish dialects and more settings in which ASR may be applied.

Table 2. An overview of the most recent applications of KNLP.

Ref	Objective	Technique(s)	Kurdish Dialect	Application (Field)	Dataset	Tools	Results and Accuracy	Limitations
[35], 2021	Fake news detection	Machine learning models and using TD-IDF	Sorani	Dialectology	KurdFake corpus	python	SVM 88,71% Random Forest 79.08%	Detected and answered a few questions.
[45], 2020	Next word(s) suggestion	N-gram model, Stupid-BackOff algorithm	Sorani and Kurmanji	Dialectology	websites and PDF books	R-program and R-studio	96.3%	Untested on different corpus domains
[46] 2019	Recognize simple sentences	CMUSphinx	Sorani	Speech recognition	BD-4SK-ASR	Java	Not Mentioned	Relatively a small sized corpus
[38], 2018	Translation	Rule-based Machine translation	Sorani	Machine translation	Simple sentence, Complex sentence, proverbs, Idioms and Phrases	Apertium platform	Not Mentioned	Translating simple phrases
[36], 2017	Dialect recognition	Phonetic and a style-based features	Sorani, Kurmanj and hawrami	Dialectology	Kurdish TV broadcasts	python	Not Mentioned	Dialects were not compared.
[41], 2017	Transliterate	Word collocation, bi-dialectal dictionary	Sorani and Kurmanj	Machine translation	Kurdish media and universities websites	python	71% for Kurmanji and 79% for Sorani	Used bi-dialectal dictionaries instead of a parallel corpus.

[43], 2016	Detecting sub dialectal differences	Classification methods and N-gram model	Sorani	Dialectology	Kurdish media	Java	96%	Lack of Kurmanji Kurdish dialects.
[37], 2016	Identify dialects	Supervised machine learning (SVM)	Sorani and Kurmanji	Dialectology	Several Kurdish media websites	Java	Not Mentioned	Better used stemmer for weighted list and categorization.
[40], 2012	Text categorization	N-gram model	Sorani	Dialectology	Government and Kurdish newspapers websites	Java	Not Mentioned	Use of few documents as they did not match the terms.
[44], 2011	Text-to-Speech	Concatenative Synthesis method	Sorani	Speech recognition	Group of volunteer listeners	Java	Not Mentioned	Speech-to-Text is not addressed

As was shown in the literature review above, variety research methods and approaches have been employed to solve a wide range of issues. The following research used Kurdish NLP applications, ranging from (2011) to the present (2021). A summary of the latest applications of Kurdish language NLP studies is shown in Table 2, As outlined in Table, a series of activities may be addressed using a variety of techniques, including machine learning models, N-gram models, rule-based machine translation, and classification approaches for both Kurdish dialect Sorani and Kurmanji, for the purpose of evaluating the approach performance, the usage of multiple datasets based on the aim is required. Overall, the author concentrated on three distinct fields: dialectology, speech recognition, and machine translation.

The authors in [35], [36], [40], [43], [45] have utilized an N-gram model for dialectology and a verity dataset, and they have several shortcomings, such as: in article [35] there are still many open questions and gaps that have not been addressed. In article [45] it is not tested on a variety of sports, economics, technology, and medicine-related books. There was no mention of the degree of resemblance across dialects in article [36]. In article [43] it covered just a few Sorani dialects as opposed to all Kurdish dialects such as Kurmanj. In article [40] few materials were utilized since they did not stem the phrases. The authors of [38], [41] used Rule-based and dictionary-based translation, both of which have a variety of drawbacks. The proposed approach in article [38] can only be used to translate a few short lines and also Instead of a parallel corpus, bi-dialectal dictionaries were used in article [41]. The authors of [44], [46] employed concatenative Synthesis and CMUSphinx for speech recognition with many constraints, such as the absence of a Kurdish Speech-to-Text reverse technique and the absence of a large corpus for ASR.

5 Kurdish Language Challenges

Developing KNLP seems promising and interesting results have been proposed for a limited number of NLP steps, however it is not free of challenges. The literature has already listed some notable limitations in the way of KNLP, challenges categorized into five groups. While the first two groups are concerned with the diversity aspect of the Kurdish language, the third and fourth highlight the processing difficulties and the last one examines the depth of resource-scarcity for Kurdish. The limitations and challenges in KNLP are [8]:

5.1 Dialect Diversity

The primary difficulty in analyzing Kurdish texts is the dialect variety. In this study, we focus on the two most prominent Kurdish dialects, Kurmanji and Sorani, with respect to their number of speakers and level of standardization. Nearly three-quarters of native Kurdish speakers [4] are accounted for by this combination. Some of the major variations in morphology [4], [47]:

- In regards to gender assignment, the Kurmanji dialect is more restrictive than the Sorani dialect, despite the fact that there are exceptions even in the Kurmanji dialect.

- Unlike Kurmanji, which has a strong conservatism of case opposition (oblique and absolute), the Sorani dialect does not. There are complete ergative transitive verbs in the past tense of Kurmanji, but Sorani does not have them.
- Passive and causative are generated differently in each dialect as well; in Sorani, the verb undergoes a morphological change to create them, but in Kurmanji, they are formed by helper verbs.

5.2 Script Diversity

Due to geopolitical considerations [8] the dialects of Sorani and Kurmanji are written in two completely different alphabet systems and lack of standardization of Kurdish writing. The common alphabets are shown in Fig. 3 with the mappings among them.

	Arabic based	Latin based
1	ێ/ئ	I
2	و	U/W
2	ی	Y/î
3	ه	E/H

(a) One-to-Two Mappings

	Arabic based	Latin based
1	ڕ	(RR)
2	ڵ	-
3	ع	(E)
4	غ	(X)
5	ح	(H)

(b) One-to-Zero Mappings

	Arabic based	Latin based		Arabic based	Latin based
1	ا	A	13	ن	N
2	ب	B	14	ۆ	O
3	ج	C	15	پ	P
4	چ	Ç	16	ق	Q
5	د	D	17	ر	R
6	ێ	Ê	18	س	S
7	ف	F	19	ش	Ş
8	گ	G	20	ت	T
9	ژ	J	21	وو	Û
10	ک	K	22	ڤ	V
11	ل	L	23	خ	X
12	م	M	24	ز	Z

(c) One-to-one Mapping

Fig. 3. (a, b, c): The two standard Kurdish alphabets

5.3 Per-processing

The process of pre-processing in Kurdish is extremely challenging for a variety of reasons, some of which will be mentioned below.

- A variety of Kurdish Unicode keyboards make use of characters that aren't native to Kurdish Unicode [48]. This is due to the fact that some Kurdish letters share their typography with letters in other Arabic-based scripts, such as Persian and Arabic as shown in Fig. 4 [49].
- Conjunctive ‘و‘meaning “and” as an individual word, Kurdish language standards define the conjunctive ‘و ‘as a separate word. However, some Kurdish publishers and academicians are unaware of this, thus they add conjunctive ‘و ‘to a proceeding word, generally one that ends in a vowel [50], [51].
- There are two distinct ‘r’s that are both written and spoken differently. Many Kurdish scholars make the mistake of writing ‘ر’ at the beginning of words when they should be writing ‘ڕ’, in Sorani Kurdish, the letter ‘ر’ never comes as the first letter of words [49].

Unicode Value	Latin based Char.	Arabic based Char.	Unicode Value
0048	H	ه	06BE
0068	h		
0049	I	-	-
0069	i		
0055	U	ێ+وو	0648
0075	u		
0057	W	و	0648
0077	w		
0059	Y	ی	06CC
0079	y		

(a) From Latin-based to Arabic-based

Unicode Value	Arabic based Char.	Latin based Char.	Unicode Value
00648	و	U	0055
		u	0075
		W	0057
		w	0077
06BE	ه	H	0048
		h	0068
		E	0045
		e	00EA
06CC	ی	Î	00CE
		î	00EE
		Y	0059
		y	0079

(b) From Arabic-based to Latin-based

Unicode Value	Arabic based Char.	Latin based Char.	Latin based Char. (approx.)	Latin based Char.
0626	ئ	-	-	-
062D	ح	-	H	0048
			h	0068
0695	ڕ	-	R+R	0052
			r+r	0072
0639	ع	-	E	0045
			e	00EA
063A	غ	-	X	0058
			x	0078
06B5	ڵ	-	L+L	004C
			l+l	006C

(c) From Arabic-based to Latin-based (Approx.)

Fig. 4. (a, b, c): Simple mappings between Arabic-based and Latin-based Kurdish Alphabets

5.4 Segmentation and Tokenization

The technique of finding the boundaries of text components like sentences, phrases, and words is known as segmentation. Because short vowels are clear expressed in Kurdish writing systems [8], this procedure is simpler than in Persian and Arabic. Short vowels have a large role in creating ambiguity in the Arabic language, thus word meaning disambiguation, homograph resolution, and part-of-speech identification become quite challenging [52]. Despite the addition of short vowels, the Arabic-based Kurdish script has two shortcomings acquired from the Arabic writing system [48]:

- Due to the absence of capitalization in the Arabic script, sentence boundaries and Named Entities are harder to distinguish.
- Space is not a deterministic boundary delimiter or sign [48]. It can be found in a word or between words, or it may come in the middle of two words. In Persian [53] and Urdu [54], there are several ideas for resolving this issue.

5.5 Lack of Annotated corpus

The Kurdish language is an under-resourced language on the web with just raw text available as a linguistic resource. Larger-scale and more dependable corpora do not yet exist for Kurdish. Even more challenging is that there is no gold-standard dataset to assist in processing Kurdish. The variety of dictionaries accessible for the Kurdish annotated corpus and the lack of big datasets are two shortcomings [51]. Developing a huge corpus of raw text from data that is readily available on the Internet may be utilized for information retrieval applications, for example: While the data obtained from the Internet does present certain issues, the ambiguity of characters must be resolved in the course of normalizing raw text [49]. The majority of Kurdish written materials have not been digitized, and the ones that are available are either not accessible or not completely convertible. The lack of a uniform orthography and the inability to use Unicode keyboards contribute to a lack of standardization in text processing in Kurdish [4].

6 Discussion

After presenting the challenges related to the nature of the Kurdish language, this section presents the challenges and drawbacks related to the proposed approaches and used techniques along with the challenges that appear when applying KNLP into different types of applications. As far as tokenization, it is quite challenging to detect compound verbs [20], besides tense, aspect, person, and mood, verbs are also inflected according to the patient. Also, the proposed approaches in this regard are mostly rule-based and relying on *Punkt* Sentence Tokenizer. For the POS tagging step of KNLP, despite promising results the proposed literature either used a relatively small-size corpus for training the proposed models [23] or did ignore the syntactical features which may modify morphological forms in a sentence.

The literature on lemmatization and stemming needs a trial of advanced ones such as deep neural networks will be worth attempting for the same used dataset instead of use of traditional classification and rule-based algorithms with more [1], [19], [25]. Also the lack of solving the problem with some of the systematic stemming errors (e.g., over-stemming and mishandling of named entities) [26], [27] and the lack of a significant number of Kurdish lemmas will be required [28]. To address the NER recognition problem, the suggested experiments employed condensed name and feature lists that excluded the great majority of Kurdish names, yielding less than encouraging results [29]-[31].

As far as applying KNLP in real-world applications, the literature used KNLP tools with classical text mining techniques such as TF-IDF, LBP and n-gram rather than more advanced ones [35], [36], [40] or not fully using KNLP phases [37]. Also the applications have a narrow output such as only translating simple sentences, compound sentences, phrases, and idioms from English to Kurdish [38], [39], [44] or use of a small-size input corpus [40]-[42], [45], [46].

According to our survey, Sorani Kurdish POS-tagging will be the future direction for Kurdish natural language processing community. While some progress has been made in KNLP's POS tagging phase, it has only been achieved with a limited number papers despite promising results in [10], [23]. The proposed literature for the Sorani dialect either used a rule-based approach with only four tag categories as a base (there are only four categories: verbs, nouns, adjectives and adverbs) or ignored the syntactical features that may modify morphological forms in a sentence [10]. The suggested models for the Kurmanji dialect were trained using a corpus of relatively limited size for the Kurmanji dialect [23].

7 Conclusion

The increase in the volume of machine-readable Kurdish web content has notably caused a rising need for the existence of the Kurdish Natural Language Processing (KNLP) pipeline. In the field of artificial intelligence, NLP has helped in developing intelligent applications which are capable of both human languages understanding and producing, to a good degree, by incorporating multiple fields such as computer science and computational linguistics. On the other hand, the new field of Kurdish NLP (KNLP) is presenting noteworthy projects, but they are not enough to support industry level semantic web, text mining and information retrieval applications for Kurdish language text. This review paper addressed the KNLP work in the literature and its applications and listed the state-of-the-art progress in the field of KNLP and its applications beside the language-related challenges, and following is the conclusion:

- The KNLP is in its middle stage of development, so researchers can focus now on its NLP pipeline from the step of POS tagging and on.
- As far as applications of KNLP are concerned, since the KNLP pipeline is immature, its applications have also not yet enabled commercial-level text mining and semantic web systems.
- From a linguistics perspective, the Kurdish language has its own challenges when it comes to developing machine models for understanding and producing it through an NLP pipeline. Such challenges are dialect diversity, script diversity, and a lack of rich annotated corpora. However, these challenges can be overcome by the development of a standard Kurdish language and dialect for the Kurdish language content on the web, for example. Also populating the awareness of research on developing a fully KNLP pipeline toward enabling efficient text mining and semantic web applications for the Kurdish content on the web.

References

[1] S. Ahmadi, "KLPT–Kurdish Language Processing Toolkit," *in Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 2020, pp. 72-84.doi: 10.18653/v1/2020.nlposs-1.11.

[2] D. H. Maulud, S. R. Zeebaree, K. Jacksi, M. A. M. Sadeeq, and K. H. Sharif, "State of art for semantic analysis of natural language processing," *Qubahan Academic Journal*, vol. 1, pp. 21-28, 2021. doi: 10.48161/qaj.v1n2a44.

[3] D. H. Maulud, S. Y. Ameen, N. Omar, S. F. Kak, Z. N. Rashid, H. M. Yasin, et al., "Review on Natural Language Processing Based on Different Techniques," *Asian Journal of Research in Computer Science*, pp. 1-17, 2021. doi: 10.9734/ajrcos/2021/v10i130231

[4] G. Haig and Y. Matras, "Kurdish linguistics: a brief overview," *STUF-Language Typology and Universals*, vol. 55, pp. 3-14, 2002. doi: 10.1524/stuf.2002.55.1.3.

[5] A. Hassanpour, J. Sheyholislami, and T. Skutnabb-Kangas, "Introduction. Kurdish: Linguicide, resistance and hope," *International Journal of the Sociology of Language*, vol. 2012, pp. 1-18, 2012. doi: 10.1515/ijsl-2012-0047.

[6] H. Veisi, M. MohammadAmini, and H. Hosseini, "Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus," *Digital Scholarship in the Humanities*, vol. 35, pp. 176-193, 2020. doi: 10.1093/llc/fqy074.

[7] G. Tavadze, "Spreading of the Kurdish Language Dialects and Writing Systems Used in the Middle East," Bull. Georg. Natl. Acad. Sci, vol. 13, 2019. Available at: http://science.org.ge/bnas/t13-n1/24_Tavadze.pdf.

[8] K. S. Esmaili, "Challenges in Kurdish text processing," arXiv preprint arXiv:1212.0074, 2012. doi: 10.48550/arXiv.1212.0074.

[9] S. Ahmadi, "A rule-based Kurdish text transliteration system," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 18, pp. 1-8, 2019. doi: 10.1145/3278623.

[10] S. Ahmadi and H. Hassani, "Towards Finite-State Morphology of Kurdish," arXiv preprint arXiv:2005.10652, 2020. doi: 10.48550/arXiv.2005.10652.

[11] T. Jugel, "On the linguistic history of Kurdish," *Kurdish Studies*, vol. 2, pp. 123-142, 2014. doi: 10.33182/ks.v2i2.398

[12] Y. Matras, "Revisiting Kurdish dialect geography: Findings from the Manchester Database," *Current issues in Kurdish linguistics*, vol. 1, p. 225, 2019. doi: 10.20378/irb-56764.

[13] K. S. Esmaili and S. Salavati, "Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison," *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol.2, pp. 300-305, 2013. Available at: https://www.researchgate.net/publication/270877570_Sorani_Kurdish_versus_Kurmanji_Kurdish_An_Empirical_Comparison.

[14] G. Haig, "On the interaction of morphological and syntactic ergativity: Lessons from Kurdish," Lingua, vol. 105, pp. 149-173, 1998. doi: 10.1016/S0024-3841(98)00014-X.

[15] Y. Karimi, "On the syntax of ergativity ın Kurdish," Poznan Studies in Contemporary Linguistics, vol. 50, pp. 231-271, 2014. doi: 10.1515/psicl-2014-0016.

[16] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," Journal of Applied Science and Technology Trends, vol. 1, pp. 140-147, 2020. doi: 10.38094/jastt1457.

[17] A. Rajput, "Natural language processing, sentiment analysis, and clinical analytics," in Innovation in Health Informatics, ed: Elsevier, 2020, pp. 79-97. doi: 10.1016/B978-0-12-819043-2.00003-4.

[18] A. Rokade, B. Patil, S. Rajani, S. Revandkar, and R. Shedge, "Automated grading system using natural language processing," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1123-1127. doi: 10.1109/ICICCT.2018.8473170.

[19] A. M. Mustafa and T. A. Rashid, "Kurdish stemmer pre-processing steps for improving information retrieval," Journal of Information Science, vol. 44, pp. 15-27, 2018. doi: 10.1177/0165551516683617.

[20] S. Ahmadi, "A Tokenization System for the Kurdish Language," in Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, 2020, pp. 114-127. Available at: https://aclanthology.org/2020.vardial-1.11.

[21] Abdulrahman, Roshna Omer, Hossein Hassani, and Sina Ahmadi. "Developing a fine-grained corpus for a less-resourced language: the case of Kurdish." arXiv preprint arXiv:1909.11467 (2019). doi: 10.48550/arXiv.1909.11467.

[22] Esmaili, Kyumars Sheykh, and Shahin Salavati. "Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013. Available at: https://www.researchgate.net/publication/270877570_Sorani_Kurdish_versus_Kurmanji_Kurdish_An_Empirical_Comparison.

[23] G. Walther, B. Sagot, and K. Fort, "Fast development of basic NLP tools: Towards a lexicon and a POS tagger for Kurmanji Kurdish," in International conference on lexis and grammar, 2010, p. 0. Available at: https://inria.hal.science/hal-00510999/.

[24] P. Denis and B. Sagot, "Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging," Language resources and evaluation, vol. 46, pp. 721-736, 2012. doi: 10.1007/s10579-012-9193-0.

[25] A. M. Saeed, T. A. Rashid, A. M. Mustafa, R. A. A.-R. Agha, A. S. Shamsaldin, and N. K. Al-Salihi, "An evaluation of Reber stemmer with longest match stemmer technique in Kurdish Sorani text classification," Iran Journal of Computer Science, vol. 1, pp. 99-107, 2018. doi: 10.1007/s42044-018-0007-4.

[26] S. Salavati, K. S. Esmaili, and F. Akhlaghian, "Stemming for Kurdish information retrieval," in Asia Information Retrieval Symposium, 2013, pp. 272-283. doi: 10.1007/978-3-642-45068-6_24.

[27] J. B. Lovins, "Development of a stemming algorithm," Mech. Transl. Comput. Linguistics, vol. 11, pp. 22-31, 1968. Available at: https://apps.dtic.mil/sti/citations/AD0735504.

[28] S. Salavati and S. Ahmadi, "Building a Lemmatizer and a Spell-checker for Sorani Kurdish," arXiv preprint arXiv:1809.10763, 2018. doi: 10.48550/arXiv.1809.10763.

[29] H. Hassani, "A method for proper noun extraction in Kurdish," in 6th Symposium on Languages, Applications and Technologies (SLATE 2017), 2017. Available at: https://drops.dagstuhl.de/opus/volltexte/2017/7952/pdf/OASIcs-SLATE-2017-19.pdf.

[30] P. Littell, K. Goyal, D. R. Mortensen, A. N. Little, C. Dyer, and L. Levin, "Named entity recognition for linguistic rapid response in low-resource languages: Sorani kurdish and tajik," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 998-1006. Available at: https://aclanthology.org/C16-1095.

[31] H. Simpson, C. Cieri, K. Maeda, K. Baker, and B. Onyshkevych, "Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources," Collaboration: interoperability between people in the creation of language resources for less-resourced languages, vol. 7, 2008. Available at: https://www.isca-speech.org/archive/pdfs/saltmil_2008/simpson08_saltmil.pdf.

[32] V. Keselj, "Speech and Language Processing Daniel Jurafsky and James H. Martin," ed: Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009. Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

[33] E. Agirre and P. Edmonds, Word sense disambiguation: Algorithms and applications vol. 33: Springer Science & Business Media, 2007. doi: 10.1007/978-1-4020-4809-8.

[34] Roberto Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, No.:10, pp 1–69, doi: 10.1145/1459352.1459355.

[35] R. Azad, B. Mohammed, R. Mahmud, L. Zrar, and S. Sdiqa, "Fake News Detection in low-resourced languages “Kurdish language” using Machine learning algorithms," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, pp. 4219-4225, 2021. Available at: https://turcomat.org/index.php/turkbilmat/article/view/8393.

[36] A. K. Al-Talabani, Z. K. Abdul, and A. A. Ameen, "Kurdish Dialects and Neighbor Languages Automatic Recognition," ARO-The Scientific Journal of Koya University, vol. 5, pp. 20-23, 2017. doi: 10.14500/aro.10167.

[37] H. Hassani and D. Medjedovic, "Automatic Kurdish dialects identification," Computer Science & Information Technology, vol. 6, pp. 61-78, 2016. doi: 10.5121/csit.2016.60307.

[38] K. M. Kaka-Khan, "English to Kurdish Rule-based Machine Translation System," UHD Journal of Science and Technology, 2018. doi: 10.21928/uhdjst.v2n2y2018.pp32-39.

[39] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, et al., "Apertium: a free/open-source platform for rule-based machine translation," Machine translation, vol. 25, pp. 127-144, 2011. doi: 10.1007/s10590-011-9090-0.

[40] F. Mohammed, L. Zakaria, N. Omar, and M. Albared, "Automatic Kurdish SORANi text categorization using N-gram based model," in 2012 International Conference on Computer & Information Science (ICCIS), 2012, pp. 392-395. doi: 10.1109/ICCISci.2012.6297277.

[41] H. Hassani, "Kurdish interdialect machine translation," in Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial), 2017, pp. 63-72. doi: 10.18653/v1/W17-1208.

[42] X. Zhang, "Dialect MT: a case study between Cantonese and Mandarin," in COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics, 1998. doi: 10.3115/980432.980807.

[43] S. Malmasi, "Subdialectal differences in sorani kurdish," in Proceedings of the third workshop on nlp for similar languages, varieties and dialects (vardial3), 2016, pp. 89-96. Available at: https://aclanthology.org/W16-4812.pdf.

[44] H. Hassani and R. Kareem, "Kurdish text to speech (KTTS)," in Tenth International Workshop on Internationalisation of Products and Systems, 2011, pp. 79-89. Available at: https://www.researchgate.net/publication/295092948_Kurdish_Text_to_Speech_KTTS.

[45] H. K. Hamarashid, S. A. Saeed, and T. A. Rashid, "Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji," Neural Computing and Applications, vol. 33, pp. 4547-4566, 2021. doi: 10.1007/s00521-020-05245-3.

[46] A. Qader and H. Hassani, "Kurdish (sorani) speech to text: Presenting an experimental dataset," arXiv preprint arXiv:1911.13087, 2019. doi: 10.48550/arXiv.1911.13087.

[47] D. N. MacKenzie, Kurdish Dialect, Studies 1 vol. 2: Oxford University Press, 1962. Available at: google books.

[48] M. Shamsfard, "Challenges and open problems in Persian text processing," Proceedings of LTC, vol. 11, 2011. Available at: https://www.academia.edu/3457856/Challenges_and_open_problems_in_persian_text_processing.

[49] S. Jaf, "A simple approach to unify ambiguously encoded Kurdish characters," in Proceedings of the International Conference Computational Linguistics in Bulgaria (CLIB 2016). 2016, pp. 86-94. Available at: https://dro.dur.ac.uk/19597/.

[50] G. Gautier, "Building a Kurdish language corpus: an overview of the technical problems," Proceedings of ICEMCO, 1998. doi: 10.1007/978-3-642-45068-6_24.

[51] G. Walther and B. Sagot, "Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish," in Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop), 2010. Available at: https://hal.inria.fr/inria-00521238/en.

[52] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, pp. 1-22, 2009. doi: 10.1145/1644879.1644881.

[53] M. Shamsfard, H. S. Jafari, and M. Ilbeygi, "STeP-1: A Set of Fundamental Tools for Persian Text Processing," in LREC, 2010. Available at: https://aclanthology.org/L10-1557/.

[54] Z. Rehman, W. Anwar, and U. I. Bajwa, "Challenges in Urdu text tokenization and sentence boundary disambiguation," in Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP), 2011, pp. 40-45. Available at: https://aclanthology.org/W11-3007.pdf.