Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 11, No. 1, pp. 127– 133, January-March 2023

journals.uoz.edu.krd

p-ISSN: 2663-628X
e-ISSN: 2663-6298

THE KURDISH LANGUAGE CORPUS: STATE OF THE ART

Media Azzat [a,*], Karwan Jacksi [a], Ismael Ali1 [a]
Department of Computer Science, Faculty of Science, University of Zakho, Kurdistan Region - Iraq
(media.azzat; Karwan.jacksi; ismael.Ali) @uoz.edu.krd.

Received: *22 Jan., 2023 /* ***Accepted:*** *16 Feb., 2023 /* ***Published:*** *20 Feb., 2023* ***https://doi.org/10.25271/sjuoz.2022.11.1.1123***

ABSTRACT:

The notable growth of the digital communities and different online news streams led to the growing availability of online natural language content. However not all natural languages have the enough attention of being made readable and comprehendible to machines. Among these less resourced and paid attention languages is the Kurdish language. Creating the machine-readable text is the first step toward applications of text mining and semantic web, such as translation, information retrieval and recommendation systems. With the de facto challenges in the Kurdish language, such as the scarcity of linguistic sources and not having unified orthography rules, this language has a lack of the language processing tools. However, to overcome the mentioned challenges and enable intelligent applications the well organized and annotated Kurdish text corpora is needed. This review paper investigates the available textual corpora in the Kurdish language and its dialects and then determined challenges are discussed, open problems are listed and future directions suggested.

KEYWORDS: Kurdish language, Text Corpus, Text Mining, Natural Language Processing.

1. INTRODUCTION

The Kurdish language is an Indo-European language, it has multi dialects, several scripts, and owns its own special grammatical system and rich vocabulary, which is spoken by approximately 30 million people in various countries, mainly in Iraq, Syria, Turkey, Armenia, Iran, and Azerbaijan [1]. The corpus indicates a set of data as an essential language resource, whereas it contains sample texts of a language. There is a considerable number of tokens and sentences, a corpus contains several word forms. It is beneficial in the linguistic analysis of a language and used in a variety of the NLP applications, such as morphology, syntax, semantics, and pragmatics [2]. There is a notable lack of resources for the machine-readable Kurdish language corpora, in both raw and annotated forms. The reasons are mainly the lack of standard orthography, and also the text of Kurdish suffers from different details in orthographic issues, beside all there is multiple keyboarding forms when it comes to text typing such as the use of non-Unicode keyboards [3]. The opensource and publicly availability of Kurdish text corpora is to enable the development of Kurdish-supported NLP pipeline, therefore enabling intelligent applications that can comprehend natural language of Kurdish. This review paper investigates and studies the available Arabic-based and Latin-based corpora, in multiple Kurdish language dialects.

2. BACKGROUND THEORY

2.1. Natural Language Processing

A language can be described as a set of rules for set of symbols. The Natural Language Processing (NLP) helps enabling the computational manipulation of the natural language resources, namely the text [4]. The NLP helps assigning the meanings to the text, as it is a part of artificial intelligence; it is enabling machines to understand the words or sentences written in human languages. The NLP has approved to enable better meaningful human-computer interaction in terms of applications [5]. The Kurdish language is a less-resourced language with few computational language studies and a lack of a unified orthography [1].

The NLP techniques have rapidly become a central component in language and speech understanding systems, Figure 1. In principle, the essence of NLP is a tool that delivers transformation change [6]. The ambiguity is the main problem handled by the NLP, including in the Kurdish language. As the Kurdish language is one of the less-resourced Indo-European languages and written using different scripts, it still lacks digital text resources to enable NLP applications [7]. The Kurdish scripts lack standardized orthographies and create differences in writing words, particularly compound forms [8].

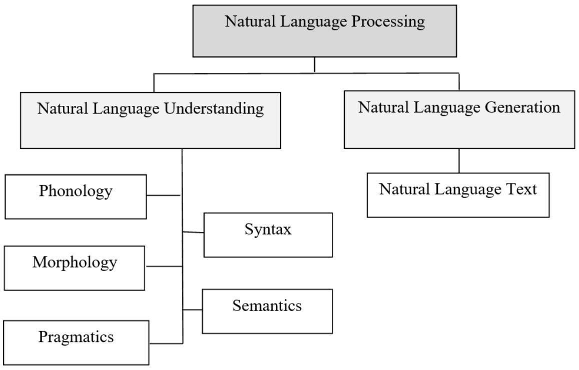

Natural Language Processing
Natural Language Understanding
Natural Language Generation
Phonology
Natural Language Text
Syntax
Morphology
Semantics
Pragmatics

Figure1: Levels of Natural Language Processing

2.2. Kurdish Language Characteristics

The tokenization is the fundamental step in the NLP pipeline, which is required for the advanced steps such as part-of-speech tagging, syntactic analysis, and machine translation [9]. The Kurdish language has prefixes and suffixes attaching to a lemma, therefore it has a complex structure in terms of morphology, especially in the Sorani dialect. The Sorani dialect lemmatization algorithms needs more accurate results [10].

* Corresponding author

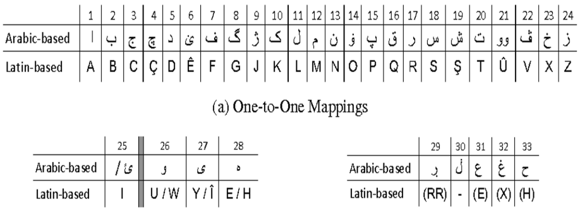
The noun in the Kurdish language is considered absolute form when it is the form without any affixes which represents a generic meaning of the word, which is the lemma provided in the dictionary. The inflection of nouns is mostly carried out with suffixes to indicate number, definiteness, indefiniteness, demonstratives, and gender. Unlike Kurmanji and Hawrami dialects, there are a few exceptions in some of the subdialects of the two latter dialects, such as the Mukryani subdialect of Sorani where nouns can have genders in specific cases. For instance, in the sentences (deçime małe Zeynebî) and (le kin Ferhadê), (Zeyneb) and (Ferhad) like as feminine and masculine proper names are inflected with the (î) and (ê) suffixes to represent the gender. The Sorani and Southern Kurdish dialects do not specify genders through morphological inflection [11]. Lemmas are used to refer to the canonical forms of the lexemes, e.g., although ((خواردن)= xwardin) (eat) and ((خواردنەوە) xwardinewe) (drink) are two distinct lexemes in Kurdish, they both have one lemma and that is xwardin [8]. There are two main types of orthography that are used to write the text in the Kurdish language, they are Arabic-based and Latin-based:

- **The Arabic-based:**

The Arabic-based orthography is predominated in regions of Iraq, Iran, and Syria. It should be noted that the Arabic-based script has a subject, object, and verb word order in the sentence, and also does not have capitalization, it is written by starting right-to-left [1] [12].

- **The Latin-based:**

The Latin-based orthography is spread and used by the Kurds in Turkey. The vowel (i) is used in grapheme in the Latin-based script, and it has some consonant letters are composed of a punctuation mark such as an apostrophe (’e), and also there are consonants and vowel letters [7] [8]. Likewise, Latin has used subject, object, and verb to create sentences, it is used for writing by direction left-to-right [1].



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

(a) One-to-One Mappings

	25	26	27	28
Arabic-based	/ ئ	و	ی	ه
Latin-based	I	U / W	Y / Î	E / H

(b) One-to-Two Mappings

	29	30	31	32	33
Arabic-based	ڕ	ڵ	ع	غ	ح
Latin-based	(RR)	-	(E)	(X)	(H)

(c) One-to-Zero Mappings

Figure 2: Two Standard of the Kurdish alphabet with Arabic-based and Latin-based

The Kurdish language is spoken by more than 30 million people in Western Asia and globally with variety range of dialects, such as Kurmanji, Bahdini, Sorani, Zaza, Gorani, Hawrami, Mukryani, Laki, and Shabaki dialects that is the least known and extremely under-documented dialect [9] [12] [11]. The following is a brief definition of main dialects in the Kurdish language:

2.2.1 The Kurmanji Dialect

The Kurmanji, Northern Kurdish, is spoken in the northern areas of Kurdistan in Iraq, Turkey, Syria, and Iran; and it is written in Latin script. Kurmanji has more speakers than Sorani, the last one has more standardized and written resources due to it being the formal dialect of Kurdish literature. The Kurmanji dialect uses gender feminine and masculine, and also it has grammatical such as phonological, lexical, and morphological [1] [13]. It is worth mentioning that in addition to Kurmanji being a dialect and also it can be used the formed with the helper verbs (*hatin*) means (“to come”) and (*dan*) means (“to give”). Like the rest of the dialects, it has four cases nominative, construct, oblique, and vocative; this dialect has postpositions and prepositions, and has combinations of these which form circumpositions [14].

2.2.2 The Sorani Dialect

Despite, the Sorani, Central Kurdish, dialect is spoken mainly in southeastern regions, including Iraq and Iran, and is mostly written in a customized version of the Arabic script. It has phonological, lexical, morphological, and sometimes semantics, The features distinguishing the Sorani dialect is no gender. The cases of possessive pronouns, definiteness markers, enclitics, and postpositions when used as suffixes are used in Sorani’s writing, [1] [13]. More, it has two tenses verbs past and present, and also has cases singular and plural, but has complex morphology [10].

2.2.3 The Badini Dialect

The Badini is nearest to the Sorani dialect, and also nearest to the Arabic language. The speakers of Badini dialect are located in the region that includes the Duhok province of Iraqi Kurdistan and the Hakkari region of southeastern Turkey. This dialect is written by one script named the Arabic-based used in the Kurdish language [4].

2.2.4 The Zazaki Dialect

The Zazaki, DimlˆI, dialect uses a Latin-based alphabet, and it is a subgroup of the Northwestern Iranian and Turkey. The speakers of the Zazaki dialect are approximately 2 million in different regions of Turkey [1] [12].

2.2.5 The Gorani Dialect

The Gorani dialect is spoken by smaller populations, it is also a subgroup of the Northwestern Iranian languages. the speaker of Gurani is approximately 300,000 speakers in the different parts of the Iranian and Iraqi Kurdistan. It is written in the Arabic-based alphabet [1] [12].

The first corpus of Kurdish language had been an initiative created back in 1998. Some scholars did estimate the first written Kurdish text were appearance in circa 1600. The first attempts to present a standard writing system for the Kurdish language began in the 1920s. Therefore in 1932 has been presented Latin script the Jeladet Ali Bedirkhan in the Kurdish language, Celadet Elî Bedirxan, also known as Bedirxan alphabet. In Iraq, scholars have presented scripts based on the Persian-Arabic. The Persian-Arabic and Latin-based scripts have been spread in different Kurdish-speaking regions [7].

Technically, there are different classes of corpora, which help in investigating probabilistic and gradient properties of a given language, and finding and discovering the interpreting cross-linguistic generalizations depending on the processing and communicative mechanisms [19].

- Parallel corpora: parallel or multilingual corpora, this type of corpus has differed in the degree of similarity between the texts in the language. Thus, which are composed aligned with the sentences or other chunks of text in more than one language, and display the highest semantic and pragmatic similarity between the components [19].
- Comparable corpora: the comparable corpora have not similar text to the parallel corpora, especially in different languages, but it represents similar text types or topics [19].
- Unified annotation: it focuses on the processing tools and unified annotating more than the similarity of texts represented in different languages [19].

2.3. Kurdish Language Corpora (KLC)

A corpus is a large well-organized collection of electronic text or a large structured set of texts. A corpus can be created from written language, spoken language or both. The sources are essentially audio recordings for spoken language, web texts, religious texts, educational texts, historical texts and etc. The annotated corpus is a corpus with POS tagging or other lexical, morphosyntactic, semantic, or pragmatic information included for building an automatic POS tagger. Although, the lack of corpora is one of the main gap in the Kurdish language processing [3] [20], creating a corpus for processing Kurdish textual data has several challenges exist that need to be addressed in this area. In this section, we review the major corpora for the Kurdish language [15]. A useful starting point for any study of this kind is to begin by defining the key terms of Annotated and Un-Annotated corpus:

- **Annotated KLC**

The Annotation corpus consists of the application of a scheme to texts, as the corpus is much more useful when annotated. For making corpora extra useful for doing linguistic research, they are usually subjected to a process known as an annotation. An example of annotating a corpus is POS tagging, in which information about each word such as parts of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags, or indicating the lemma (base) form of each word [6]. The available annotated corpora as follows:

- The Kurdish Textbooks Corpus (KTC) are containing 693K Sorani words, it is composed of 31 K-12 textbooks in the Sorani dialect [3].
- The Wergor corpus to provides a resource for Kurdish transliteration, this corpus consists of parallel transliterated texts from the two orthographies [21].
- The Tanzil corpus is a compilation of Quran translations for Sorani Kurdish, which is aligned with 11 translations in English making a total number of 92,354 parallel sentences with 3.15M words on the Sorani Kurdish side and 2.36M words on the English side [15].
- The TED corpus is used for Technology, Entertainment, and Design, The Sorani Kurdish dialect is the only Kurdish dialect for which these subtitles are translated. it is small in size approximately 2358 parallel sentences [15].
- The KurdNet–the Kurdish WordNet is a lexical-semantic resource, this is a corpus that provides translations in Sorani Kurdish. In addition to semantic relationships such as synonymy, hyponymy, and meronymy, the current version of KurdNet contains 4,663 definitions, it is directly translated from the Princeton WordNet (version 3.0) [7].
- The AsoSoft text corpus for Kurdish Sorani has 188 million tokens [1].
- Bianet is parallel news corpus containing 6,486 English-Kurmanji Kurdish and 7,390 Turkish-Kurmanji Kurdish sentences [22].

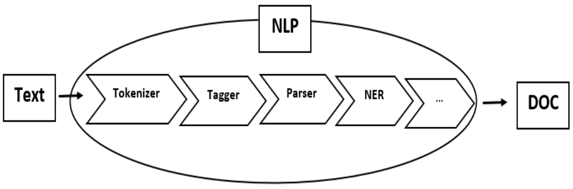


Figure3: Annotated Document

- **Un-Annotated KLC**

The Un-Annotation corpus consists of the part of text, also a corpus may contain texts in a single language or multiple languages. A Kurdish text corpus is collected from different sources, such as Web sites and PDF books. Kurdish Sorani texts are different from Kurmanji text, for instance, in terms of the morphology of the letters. Accordingly, all the texts are transformed into Kurmanji using the Kurdish [6]. In the following, we describe the available corpora of data used for un-Annotation KLC:

- The Pewan corpus are containing 18M Sorani words, and 4M Kurmanji words.
- The corpus of (2020)'s Sorani corpus containing 8.1M words.
- The Sorani Kurdish folkloric lyrics corpus containing 49K words [23].
- The raw text of the Kurmanji Wikipedia containing 3M words [8]. This corpus contains more than 500,000 words and includes words of different topics, such as social, religious, economic, etc.

3. LITERATURE REVIEW

This section presents the major existing Kurdish text corpora, their description, applications and limitations.

The work in [24] used the methods of Kurdish language classification especially for Sorani texts for detecting Iran and Iraq subdialects, including 200,000 sentences. In order to maintain a balance between short and long sentences, they used a dataset of sentences in the range of 5–55 tokens and extracted data from the news outlets, such as Rudaw, Sharpress, Sahar TV and Kurdpress. Two types of lexical surface features were used: character n-grams and word n-grams, particularly 4-grams. This corpus does not include the NER tags. The limitations in this work are to provide detailed analysis and expanding the dataset with additional data from different resources.

Authors in [14] created Kurdish Treebank, as developed syntactically to annotate corpus for the Kurmanji Kurdish, which has based on the Apertuim. This corpus was used as a CoNLL 2017 shared task on parsing universal dependencies, which were used in tokenization, parts of speech, and morphological analysis. For evaluation and testing the treebank been used three parsers Maltparser, UDPipe, and BiST with and without dictionary, and the UDPipe parser have the accuracy with a dictionary performed UAS (74.3%) and LAS (67.9%).

In [10], authors presented lemmatization and word-level error correction system that is a hybrid approach depending on the rules-based method of morphological and used n-gram modeling, which called Peyv as lemmatization, and Renus as spell-checker. The Peyv and Renus systems used the corpus called Pewan, it holds about 18M words from 115K news articles. The proposed systems have obtained 86.7% accuracy for the Peyv Lemmatizer, the Renus with 96.4% accuracy, and the correction system with 87% accuracy without lexicon. The Peyv and Renus are considered a good step system for used lemmatization, and spell-checker.

The authors in [22] paper collected and used corpus from Bianet magazine, which is an online newspaper that publishes Turkish news, that is translated from English-Kurdish and Turkish-Kurdish. It is a bilingual text at the sentence level. The corpus consisted 35,080 sentences, and 1,3 million tokens. For evaluating the system had been utilized the BLEU and chrF3 automatic evaluation metrics; and used the Multeval test for evaluating the accuracy of the models. The proposed system can be expanded to different dialects of the Kurdish language.

In [3], the authors have proposed the Kurdish Textbooks Corpus (KTC) and it is featured by the Ministry of Education of the Kurdistan Region of Iraq. It was collected from 31 K12 textbooks in the Sorani dialect, and the corpus was printed and classified into 12 educational subjects containing 693,800 tokens and 110,297 types. Additionally, converting the KTC to the Unicode, the Unicode has been used at a pre-processing stage, and the texts were normalized by replacing (ZWNJ) zero-width-non-joiner and manually verifying the orthography. but without removing punctuation and special characters. This work can be more valuable if composed of a variety of texts in the terms of the different domains of knowledge such as texts from sports, technology, and medicine.

The AsoSoft corpus is presented in [1] with the size of 188 million tokens and it has been composed of Web sites, magazines, and published books, where the text had been normalized and converted for processed text to the Text Encoding Initiative (TEI) format for standardization of data. On the other hand, it was subject annotated with six topic tags approximately 22% of the corpus to evaluate the correctness of the annotation. This corpus was applied to Zipf's law and perplexity was calculated using a statistical model of N-gram language models of the Kurdish language. It is worth mentioning that the AsoSoft corpus has created two Crawlers on the website, such as Apache Nutch and PHP Crawlers. The collected data has been between 2003 to 2015 and included fifty-two Web sites that were taken in Kurdish text. The AsoSoft has hopeful results but the corpus is needing more normalization for non-standard tokens and to correct misspelled.

The work in [23] presented Kurdish folk lyrics as a corpus that covers various Kurdish musical genres, including Beyt, Goranˆı, Bend, and Heyran in the Sorani dialect of the Kurdish language. In this corpus, authors transformed the transcription into a structure of XML according to the TEI, so the development of the corpus is carried out by transcribing folkloric songs manually from audiovisual materials. Overall, this corpus contains 49,582 tokens, and they transcribed a set of 162 songs of Kurdish folkloric. The Folkloric Lyrics corpus is important for many areas of study such as named entity recognition, relation extraction, computational musicology and coreference resolution.

Authors in [25] have used machine learning for the Sorani dialect of the Kurdish language to segment the textbook corpus - KTC, that it is written in Persian-Arabic script. The corpus was tokenized by using NLTK's word tokenizer, and contained 693,800 tokens and 110,297 types, and saved in an XML file. Finally, in this test, they used and achieved by F1 score of 91.10% and had an Error Rate of 16.32%. The high Error Rate is mainly due to the state of abbreviations as parameters in the Kurdish. The proposed corpus is considered a good step toward a more developed corpus, but it is limited to expanding and applying the NLP tools.

A corpus of Zazaki and Gorani languages presented in [12]. It depended on the news articles from different resources on several topics such as science, politics, culture, and art, and it contains over 1.6M word tokens in Zazaki and 194k word tokens in Gorani. It is worth mentioning that the Zazaki dialect uses a Latin-based alphabet, while the Gorani dialect uses an Arabic-based script. The topic of the corpus focused on cultural issues which are composed of analytical articles in humanities and provide interviews in such fields. And they selected Zazaki.net for Zazaki and Firat News Agency, which has a wider range of news in topics such as women, politics, the world, Kurdistan, science, culture, and art for Zazaki and Gorani languages. The Zazaki articles dated between 2009 and 2020 while the Gorani ones are more recent 2018 to 2020. Finally, the corpus of Zazaki and Gorani dialects was applied to Zipf's law. The corpus is collected between the two different dialects in the scripts, but the tasks of NLP pipeline were not applied completely.

The work in [15] created a machine translation system for the Sorani dialect, therefore they used three types of corpora for the Kurdish language such as the Tanzil corpus, the TED corpus, and KurdNet–the Kurdish wordnet. The corpus included 92,354 parallel sentences with 3.15M words in the Sorani Kurdish. The second corpus is the collection of subtitles from TED Talks which are a series of high-quality talks on Technology, Design, and Entertainment. The TED corpus has 2358 parallel sentences from Sorani dialects. The last corpus is the WordNet of a lexical-semantic resource that contains 4,663 definitions. There are some limitations in this work, such as not including the use of a task morphology analysis. But the system has been tested the tokenization methods for developing a neural machine translation.

The authors in [26] described an approach for retrieving potentially-alignable articles of news from websites dependent on lexical similarity and transliteration of scripts. This corpus included 12,327 translation pairs for dialects Sorani, and Kurmanji as well as including 1,797 and 650 translation pairs in English-Sorani and English-Kurmanji. The data crawling in this paper was included from the content of news websites from Firat News, Agency (ANF), BasNew (BN), and KurdPa (KP). This corpus was written in the Arabic-based alphabet for Sorani dialect, and Kurmanji dialect was written for the Latin-based script. this corpus was prepared a good translation between the two important dialects Sorani, Kurmanji and translation to English language. The author employed a parallel corpus for the Kurdish language, which consider as a less-resourced language. therefore, need more developments in Kurdish machine translation.

Sina Ahmadi in [8] has created a technique for tokenization system for two dialects of Sorani and Kurmanji of the Kurdish language. To evaluate the performance of the tokenization, the system has created a dataset of the gold standard used 100 sentences from Pewan and KTC annotated corpus for Kurmanji and Sorani respectively. Due to the limited advances in Kurdish natural language processing, this work could evaluate tokenization as a component alone, but there are some limitations to the research; including not employing tokenization of compound verbs.

Zhila Amini et.al. have presented the machine translation application, called Awta Central Kurdish-English [27]. A corpus was built for the machine translation depending on the different morphological, syntactic, and semantic levels. This corpus is containing 229,222 translation parallel pairs. The performance of this corpus system achieved 22.72 and 16.81 in BLEU scores for Kurdish-English and English-Kurdish. Finally, these models achieved a BLEU of 16.81 for English to Kurdish, and 22.72 for Kurdish to English translation. The system is considered a good step toward progressing the machine translation tools and understanding the intricacies of Central Kurdish, however, the proposed is need to include more translation pairs.

Hossein Hassani proposed a translated Bijankahn corpus in [28]. The lexicon of the POS-tagged of Sorani dialect was considerably less than the Farsi corpus. In order to achieve translations from one language to another in the lexicon of the Arabic-based, the manual POS-tagging is used. The POS-tagging is the hardest and most costly in the preparation of the lexicon, therefore the lexicon used more expansion in the use of entries, and improved results of more than 13,294 entries for Sorani.

The table 1 presents a detailed overview of the all 14 works that both proposed and also used Kurdish corpora of different Kurdish language dialects for different purposes.

Table1: An overview of Kurdish language corpus

N	Ref/Year	Corpus Size	Technique and Model	Language/ Dialect	Application	Text	Dataset	Outcome
1	[24] (2016)	The range of dataset 5-55 Token	(Language Identification Methods), (classification models (linear Support Vector Machine to perform multi-class)) and (n-grams)	Sorani (Iraq and Iran)	Discriminating Similar Language (DSL) Shared task	200,000 sentences	Iraq (Rudaw and Sharpress) Iran (Sahar TV and Kurdpress)	Automatic identification and Using n-grams to obtained 96% accuracy
2	[14] (2017)	10,260 Token	Used VISL formats, CoNLL, dependency parsing UDPipe, and a statistical model	Kurmanji	A Dependency Treebank	780 sentences	Sherlock Holmes story, Kurdish Wikipedia	For evaluated used Maltparser, UDPipe, and BiST. Presented treebank sets the 2017 CoNLL
3	[10] (2018)	18M Words	Used n-gram model, CoNLL, and rule-based	Sorani	Peyv and Renus	115K news articles	Pewan text Corpus	The Peyv Lemmatizer achieved 86.7% accuracy, the Renus had achieved 96.4% accuracy; the correction system had 87% accuracy without lexicon
4	[22] (2018)	1,3M Tokens	used Python and HTML tags, and also used the HunAlign sentence aligner	Turkish, Kurdish and English	Bianet corpus	35,080 sentences	Bianet magazine, an online newspaper	To evaluated system used WMT 2016, BLEU and chrF3.
5	[3] (2019)	693,800 Token	converted to Unicode and used Zipf's law	Sorani	Kurdish textbook corpus (KTC)	110,297 types	31 K-12 textbooks	Semi-automatic conversion
6	[1] (2019)	188M Tokens	Used Text Encoding Initiative (TEI), XML format and Zipf's law, and used machine learning algorithms, including (SVM), (SMO) algorithm, BayesNet, naïve Bayes, and decision tree.	Sorani	AsoSoft text corpus	458,000	Web sites (Published books, and magazines)	Annotated document classification, using F-measure criteria, and TF-IDF for evaluated
7	[23] (2020)	49,582 Tokens	Text Encoding Initiative (TEI), and XML format. transcription process by audiovisual materials	Sorani	Kurdish Folkloric Lyrics	162 songs	Beyt, Goranˆı, Bend, and Heyran	Provides additional linguistic information
8	[25] (2020)	693,800 Tokens	Punkt, an unsupervised machine learning method	Sorani	Kurdish textbook corpus (KTC)	110,297 types	Kurdish corpus	Segmented text to sentences, and achieved 19.35% accuracy
9	[12] (2020)	1,633,770 Zazaki Tokens 194,563 Gorani Tokens	Used Zipf's law and Separate by JSON file	Zaza and Gorani	corpus in Zazaki and Gorani	(Zazaki 102,665 and Gorani 41,454) words	News of the websites	Collecting documents and creating corpus
10	[15] (2020)	Tenzil Corpus (Kurdish 25.82 English 27.96) Ted Corpus (Kurdish 69.21 English 93.54) KurdNet Corpus (Kurdish 7.51 English 8.51) Tokens	using unsupervised techniques, the tokenization techniques are BPE, Unigram, WordPiece, and WordPunct.	Sorani	Tanzil Corpus	(Sorani 3.15M and 2.36M English) words	Quran	Develop a neural machine translation system for the Sorani dialect of Kurdish.
11	[26] (2020)	34,901-34,207 (Kurmanji - English) 11,220-11,855 (Sorani-English) 205,382-231,559 (Sorani - Kurmanji) Tokens	used HTML tags, in some cases used JSON-LD and the meta tags, and a semi-automatic manner	Sorani and Kurmanji	a Kurdish Parallel Corpus	corpus contains (12,327 Sorani-Kurmanji) (1,797 Kurmanji-English) (650 Sorani-English) translation pairs	News articles from websites	BLEU scores for the Sorani-English (17.74), Kurmanji-English (11.06), and Sorani-Kurmanji (17.08) data.
12	[8] (2020)	(9,970 Headwords for	used JSON,	Kurmanji and Sorani dialects	A Tokenization System	100 sentences	Used corpus FreeDicts, the Kurdish	For evaluated create a gold-standard dataset, and achieved accuracy 31.38 for

		Kurmanji, and 1,507 lemmata) (8,180 Headwords, and 1,513 lemmata for Sorani)	and also used unsupervised neural models such as using Hugging-Face Tokenizers and SentencePiece				Wiktionary, Wîkîferheng. and lexicon Wergor of Sorani	Kurmanji, and 30.44 for Sorani.
13	[27] (2021)	2.1 M tokens in Kurdish and 2.2 M tokens in English	Implemented and used Open NMT machine translation library, used Intertext, trained with the Adagrad optimization algorithm, and used BLEU metric ranks	Kurdish (Sorani, Kurmanji), and English	Central Kurdish-English Awta	230 k Sentences	Languages	Achieve a BLEU of 16.81 for En→Ku and 22.72 for Ku→En translations
14	[28] (2022)	Not Mentioned.	used Microsoft Bing, Libreoffice calc, and python.	Sorani	POS-tagged lexicons	13,294 entries	Bijankhan corpus	Get good results of the number, percentage, and percentile of POS Tagged in the lexicon of Sorani, which had 20,083 entries.

4. DISCUSSION

According to the reviewed literature, the researchers have used different techniques for proposing Kurdish corpora in order to overcome different problems in the area of Kurdish NLP. The table 1 is comparing the reviewed articles from 2016 to 2022. Some work focused on several dialects of the Kurdish languages, such as [22] [3] [23] [25] [12] [15] [26] [8] and focusing on the tokenization task, which is the first step from the NLP pipeline. The other set of literature of [24][10][1][27][28] worked on both tasks of tokenization, POS tagging and Name Entity Recognition to produce annotated corpora. Overall, below is summary by numbers on the reviewed literature on Kurdish based corpora:

- Eight of Sorani dialect corpora, from a variety of sources such as Iraqi Rudaw media network and Sharpress, and Irani Sahar TV and Kurdpress. Other sources from textbooks and Web sites as Pewan Corpus, Beyt, Goranˆı corpus, Quran, Kurdish corpus, and Bijankhan corpus.
- Only one Kurmanji corpus, the resources were Sherlock Holmes story, and Kurdish Wikipedia, by used A Dependency Treebank application.
- Only one Zaza and Gorani dialects corpus from News websites.
- Three corpora for Kurmanji and Sorani together.

It is worth noticing that corpora have been presented for different Kurdish dialects, however the main challenge is how to effectively crawl the web content to collect Kurdish textual data. Also, presented corpora mostly been used for the tokenization task, this is an essential model of the Kurdish languages in the NLP, but other advanced NLP steps were not been considered.

5. CHALLENGES AND OPEN PROBLEMS

The Kurdish language has been spoken by a large population in different geographical areas globally. Therefore, there is a variety of challenges in processing the Kurdish language, besides being a digitally less-resourced language. The following bellow some steps to clarify these challenges [1].

- Annotated and unannotated corpora for the Kurdish language have a notable insufficiency especially [2].
- The lack of standardization, unified grammar and raw corpora [1].
- Diversity in dialects and scripts has led to more challenges and difficulties [8].
- There are a few attempts in building multi-dialects corpora and lexicons [29].

6. CONCLUSION

The corpus plays a crucial role in expanding the area of NLP. There is a good number of efforts done for producing Kurdish text corpora, however in the terms size and dialect-diversity there is still a need of larger corpora that covers different fields and dialects. This review paper gives a generic structure and guidelines for organizing the Kurdish language corpora and methods used to address major Kurdish language dialects and corpuses; especially Sorani, Kurmanji, Zazaki, and Gorani dialects. This paper has studied the all-available Kurdish corpora in the literature, eight of Sorani dialect corpora, one Kurmanji corpus, one Zaza and Gorani dialects corpus, and three corpora for Kurmanji and Sorani together. It is notable from the reviewed literature that there is an actual need for the following categories of Kurdish corpora and future direcitons:

- Multi-language corpora, meaning text corpus with Kurdish and other languages for same text documents in the corpus.
- Annotated corpora, to enable advanced NLP and text mining applications such as ontology learning.
- A maturely developed NLP pipeline, to enable Kurdish information retrieval and semantic web.
- A professional collaboration between academia and Kurdish language linguists for developing robust Kurdish corpora, in the terms of Kurdish language syntax and semantics.

REFERENCES

[1] H. Veisi, M. MohammadAmini, and H. Hosseini, “Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus,” *Digit. Scholarsh. Humanit.*, no. June, 2019, doi: 10.1093/llc/fqy074.

[2] Z. Alyafeai, M. S. Al-shaibani, M. Ghaleb, and I. Ahmad, “Evaluating Various Tokenizers for Arabic Text Classification,” vol. 5, 2021, [Online]. Available: http://arxiv.org/abs/2106.07540.

[3] R. O. Abdulrahman, H. Hassani, and S. Ahmadi, “Developing a Fine-Grained Corpus for a Less-resourced Language: the case of Kurdish ∗,” pp. 106–109.

[4] A. Al-Talabani, Z. Abdul, and A. Ameen, “Kurdish Dialects and Neighbor Languages Automatic Recognition,” *ARO-The Sci. J. Koya Univ.*, vol. 5, no. 1, pp. 20–23, 2017, doi: 10.14500/aro.10167.

[5] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural Language Processing : State of The Art , Current Trends and Challenges Natural Language Processing : State of The Art , Current Trends and Challenges Department of Computer Science and Engineering Manav Rachna International University , Faridabad-,” *arXiv Prepr. arXiv*, no. August

2017, 2018.

[6] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait J. Sci.*, vol. 43, no. 4, pp. 95–113, 2016.

[7] S. Ahmadi, H. Hassani, and J. P. McCrae, "Towards electronic lexicography for the Kurdish language," *Proc. Electron. Lexicogr. 21st Century Conf.*, vol. 2019-Octob, pp. 881–906, 2019.

[8] S. Ahmadi, "A Tokenization System for the Kurdish Language," 2013.

[9] S. Ahmadi, "KLPT – Kurdish Language Processing Toolkit," pp. 72–84, 2020, doi: 10.18653/v1/2020.nlposs-1.11.

[10] S. Salavati and S. Ahmadi, "Building a Lemmatizer and a Spell-checker for Sorani Kurdish," *arXiv*, 2018.

[11] S. Ahmadi and H. Hassani, "Towards Finite-State Morphology of Kurdish," *arXiv*, no. Cl, 2020.

[12] S. Ahmadi, "Building a Corpus for the Zaza–Gorani Language Family," *Proc. 7th Work. NLP Similar Lang. Var. Dialects*, pp. 70–78, 2020, [Online]. Available: https://aclanthology.org/2020.vardial-1.7.

[13] P. Aliabadi, S. Salavati, M. S. Ahmadi, and K. Sheykh Esmaili, "Towards building KurdNet, the Kurdish WordNet," *GWC 2014 Proc. 7th Glob. Wordnet Conf.*, pp. 1–6, 2014.

[14] M. Gökırmak and F. Tyers, "A dependency treebank for Kurmanji Kurdish," *Proc. Fourth Int. Conf. Depend. Linguist. (Depling 2017)*, no. Depling, pp. 64–72, 2017.

[15] S. Ahmadi and M. Masoud, "Towards Machine Translation for the {K}urdish Language," *Proc. 3rd Work. Technol. MT Low Resour. Lang.*, pp. 87–98, 2020, [Online]. Available: https://aclanthology.org/2020.loresmt-1.12.

[16] H. Rouhizadeh, M. Shamsfard, V. Tajalli, and M. Rouhziadeh, "Persian-WSD-Corpus: A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation."

[17] M. Asgari-Bidhendi, B. Janfada, O. R. Roshani Talab, and B. Minaei-Bidgoli, "ParsNER-Social: A Corpus for Named Entity Recognition in Persian Social Media Texts," *J. AI Data Min.*, vol. 9, no. 2, pp. 181–192, 2021, doi: 10.22044/jadm.2020.9949.2143.

[18] A.-A. Asaad, "QuranTree.jl: A Julia Package for Quranic Arabic Corpus," *Proc. Sixth Arab. Nat. Lang. Process. Work.*, pp. 208–212, 2021, [Online]. Available: https://aclanthology.org/2021.wanlp-1.22.

[19] N. Levshina, "Corpus-based typology: Applications, challenges and some solutions," *Linguist. Typology*, pp. 1–32, 2021, doi: 10.1515/lingty-2020-0118.

[20] I. E. Onyenwe, "Developing Methods and Resources for Automated Processing of the African Language Igbo," no. April, 2017.

[21] S. Ahmadi, "A rule-based Kurdish text transliteration system," *arXiv*, vol. 1, no. 1, pp. 1–9, 2018.

[22] D. Ataman, "Bianet: A Parallel News Corpus in Turkish, Kurdish and English," *arXiv*, pp. 1–4, 2018.

[23] S. Ahmadi, H. Hassani, and K. Abedi, "A Corpus of the {S}orani {K}urdish Folkloric Lyrics," *Proc. 1st Jt. Work. Spok. Lang. Technol. Under-resourced Lang. Collab. Comput. Under-Resourced Lang.*, no. May, pp. 330–335, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.sltu-1.46.

[24] S. Malmasi, "Subdialectal Differences in Sorani Kurdish," *Proc. Third Work. NLP Similar Lang. Var. Dialects*, pp. 89–96, 2016, [Online]. Available: https://www.aclweb.org/anthology/W16-4812.

[25] H. Hassani, "Using Punkt for Sentence Segmentation in non-Latin Scripts: Experiments on Kurdish (Sorani) Texts," pp. 1–3, 2020.

[26] S. Ahmadi, H. Hassani, and D. Q. Jaff, "Leveraging Multilingual News Websites for Building a Kurdish Parallel Corpus," 2020, [Online]. Available: http://arxiv.org/abs/2010.01554.

[27] L. Informatique, "Central Kurdish Machine Translation : First Large Scale Parallel Corpus and Experiments," pp. 1–13.

[28] H. Hassani, "Part of Speech Tagging (POST) of a Low-resource Language using another Language (Developing a POS-Tagged Lexicon for Kurdish (Sorani) using a Tagged Persian (Farsi) Corpus)," 2022, [Online]. Available: http://arxiv.org/abs/2201.12793.

[29] K. S. Esmaili, "Building A Test Collection For Sorani Kurdish."

[30] Keselj, Vlado. "Speech and Language Processing" Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+ 988 pp; hardbound, ISBN 978-0-13-187321-6, $115.00." (2009): 463-466.