A Corpus of the Sorani Kurdish Folkloric Lyrics

Sina Ahmadi[1], Hossein Hassani[2], Kamaladdin Abedi[3]
[1]Insight Centre for Data Analytics, National University of Ireland Galway - Ireland
[2]University of Kurdistan Hewlêr, Kurdistan Region - Iraq
[3]Kurdistan University of Medical Sciences, Sanandaj, Iran
[1]sina.ahmadi@insight-centre.org, [2]hosseinh@ukh.edu.krd, [3]kamal.abedi@gmail.com

Abstract

Kurdish poetry and prose narratives were historically transmitted orally and less in a written form. Being an essential medium of oral narration and literature, Kurdish lyrics have had a unique attribute in becoming a vital resource for different types of studies, including Digital Humanities, Computational Folkloristics and Computational Linguistics. As an initial study of its kind for the Kurdish language, this paper presents our efforts in transcribing and collecting Kurdish folk lyrics as a corpus that covers various Kurdish musical genres, in particular *Beyt*, *Goranî*, *Bend*, and *Heyran*. We believe that this corpus contributes to Kurdish language processing in several ways, such as compensation for the lack of a long history of written text by incorporating oral literature, presenting an unexplored realm in Kurdish language processing, and assisting the initiation of Kurdish computational folkloristics. Our corpus contains 49,582 tokens in the Sorani dialect of Kurdish. The corpus is publicly available in the Text Encoding Initiative (TEI) format for non-commercial use under the CC BY-NC-SA 4.0 license at `https://github.com/KurdishBLARK/KurdishLyricsCorpus`.

Keywords: Computational Folkloristics, less-resourced languages, lyrics corpus, Kurdish

1. Introduction

Kurdish is considered a less-resourced language for which general-purpose grammars and raw internet-based corpora are the only existing resources (Hassani, 2018). While the lack of a long history of written text is considered as one of the reasons for this situation, the lack of research and activities on data collection are also counted as other reasons in this regard (Ahmadi et al., 2019).

Folkloric content play a significant role in Kurdish life as a crucial medium in communication between different Kurdish generations (Blum and Hassanpour, 1996). They are also a rich source of vocabulary, as they have mostly been traditionally transferred over generations orally and less in a written form (Kreyenbroek, 2005). A few but crucial efforts have been made to transcribe some products of the Kurdish oral literature in the beginning of the previous century by both western and eastern scholars (Rasul, 1999; Salimi, 2015; Mikailee, 2015). Because these transcripts are mainly available in hard copy and not electronic forms, they are not suitable for computational processes. On the other hand, the lack of optical character recognition systems for Kurdish prevents the automatic conversion of these resources into text formats.

Despite the limited number of resources for Kurdish, there have been various studies to create new corpora. Esmaili et al. (2013) present *Pewan*, a general-purpose corpus based on the news articles in Sorani and Kurmanji dialects of Kurdish. Similarly, Ataman (2018) presents a parallel corpus containing Kurmanji Kurdish news articles. With a particular focus on automatic identification of subdialects, Malmasi (2016) creates a corpus using articles from news sources. In the most recent attempt, Abdulrahman et al. (2019) present the Kurdish Textbooks Corpus (KTC), which is composed of 31 K-12 textbooks in Sorani dialect. Unlike previous resources which are based on news articles, the latter is more domain-specific. However, none of these resources deals with oral material and Kurdish folkloric heritage.

In this paper, we present a corpus of folkloric lyrics and songs in Sorani Kurdish containing 12, 8, 141, and 1 item respectively for four musical genres, namely *Bend*, *Beyt*, *Goranî*, and *Heyran*. The development of the corpus is carried out by transcribing folkloric songs manually from audiovisual materials and transforming the transcription into a structured format in XML according to the Text Encoding Initiative (TEI) (Ide and Véronis, 1995). Moreover, our project could be considered as an initiative to mobilize the Kurdish community to provide further documentations for the Kurdish oral literature.

This corpus can serve various aspects of natural language processing (NLP) for the Kurdish language. While it enriches the diversity of the available datasets and corpora, it also adds a set of folkloric vocabulary which could not be found in the prose and non-poetic Kurdish writing. Furthermore, the computational folkloristics (Abello et al., 2012), which has not been addressed in the context of Kurdish studies yet, can also benefit from the result of this research. As the collected songs are performed by different local singers, this can provide further insights into the subdialectal variations of Sorani Kurdish and therefore, will be beneficial to speech recognition tasks.

The rest of this paper is organized as follows. Section 2 provides an overview of the Kurdish folklore and presents the major types of Kurdish lyrics emphasizing on those that are presented in our corpus. In Section 3, we summarize what has been done with respect to the Kurdish folklore. Section 4 presents the corpus and illustrates some statistics about it. The evaluation of the corpus is given in Section 5. Finally, Section 6 concludes the paper.

2. Kurdish Folklore

The Kurdish folklore has been addressed as the major pillar of the Kurdish literature by eastern and western schol-

Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020) , pages 330–335
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

ars (Salimi, 2015; Allison, 2001; Abubakir, 2016). Traditionally, this folklore is transmitted orally. They have been influencer of and influenced by other surrounding cultures and folklores (Leezenberg and others, 2011; Rasul, 1999).

Given the diversity of dialects of the Kurdish language, there are many types and genres which are specific to each dialect. Similarly, the content of the transmitted songs might not be identical among these dialects. Furthermore, such a diversity brings a different terminology with itself which might not be similar in all dialects. For instance, the individuals who perform songs are called by different terms, such as *Dengbêj* (bard), *Stranbêj* (minstrel), and *Çirokbêj* (storyteller) in Kurmanji (Broughton et al., 2006) and, *Goranîbêj*, *Heyranbêj*, and *Beytbêj* (and *xoşxwan*) in Sorani. While *Çirokbêj* and *Dengbêj* are used interchangeably in some contexts, they refer to different types of performing (Bocheńska, 2014). In Kurmanji speaking areas, *Dengbêj* is used in a broader context as a person who sings different types of music and also plays certain instruments while singing the song (Reigle, 2014).

According to Mikailee (2015), there has not been significant academic research on the Kurdish lyrics. A survey over the existing literature indicates that there is not a common categorization for the Kurdish lyrics and further discussions about the origin of the lyrics have been ongoing among scholars (Rasul, 1999; Hassanpour, 2005). Hassanpour (2005) discusses the multi-root nature of Kurdish songs. Moreover, a common opinion states that the Kurdish lyrics have been influenced by Turkish, Arabic, Azeri, Persian, and Armenian music during a long interconnection among these ethnics (Rasul, 1999; Leezenberg and others, 2011; Hassanpour, 2005).

Given the diversity of the Kurdish lyrics in form and genres, we only focus on four types of Sorani Kurdish folkloric songs, namely, *Beyt*, *Bend*, *Goranî* and *Heyran*. A few examples of these types are illustrated in Figure 1 for comparison.

2.1. Bend

Bend is a genre of Kurdish secular narrative recital song which is performed by *bendbêj* or *şayîyer*, commonly at rural gatherings and weddings. There is no evidence to indicate when Bend dates back in the history, but a strong element of praise and adoration as one of its most important components, and also a rich structure full of love, village lifestyle, farming work, nature description, local mystics, local lords, rebellions, and warfare stories guide us to assume that it may return to where the first Kurdish local social and political power was formed (Hamelink, 2016; Brenneman, 2016). Another important feature of Bends is the improvisation element which has been evolved over time, dealing with important political and especially social issues of the day. Recently an element of nationalism has been added to Bend, making it much powerful and widespread all over the Sorani-speaking regions and sometimes even in the regions which speak other dialects such as Southern Kurdish and Kurmanji (Christensen, 2007).

2.2. Beyt

Beyt is a term in Sorani dialect for a type of lyric which is usually a long piece of work based on different subjects, such as historical, mythical, legendary, and love figures and events (Merati, 2015; Salimi, 2015). Beyts have different contexts, such as epics, historical battles, mythical tales, fables, and tragic love stories (Rasul, 1999; Sharifi, 2005; Barzegar Khaleghi, 2009; Mikailee, 2015). In some Sorani speaking areas, the term Bend is used interchangeably along with Beyt. However, Bend is usually used with a more popular content . *Beytbêj*, literally meaning Beyt sayer, recites beyt in gatherings (Barzegar Khaleghi, 2009). Although Beyts are poetic, they do not follow any particular standard for their form or size (Barzegar Khaleghi, 2009).
The transcription of Beyts in Sorani dates back to the 1900s (Rasul, 1999; Sharifi, 2005). 17 Beyts were transcribed around 1905, which were translated into Sorani Kurdish in 1975 (Rasul, 1999; Sharifi, 2005; Mikailee, 2015). From 1950s onward, other transcriptions started to appear (Mikailee, 2015). According to Sharifi (2005) and Mikailee (2015) during 1960s major transcripts were presented in Sorani Kurdish. In some cases, these transcripts were provided along with the translation into other languages, for instance, Persian (Sharifi, 2005). The transcripts by Qader Fattahi Qazi (also spelled as Ghader Fattahi Ghazi) (Sharifi, 2005; Mikailee, 2015) are examples of the efforts in this area which are also one of the major sources of the Beyts section in our corpus.

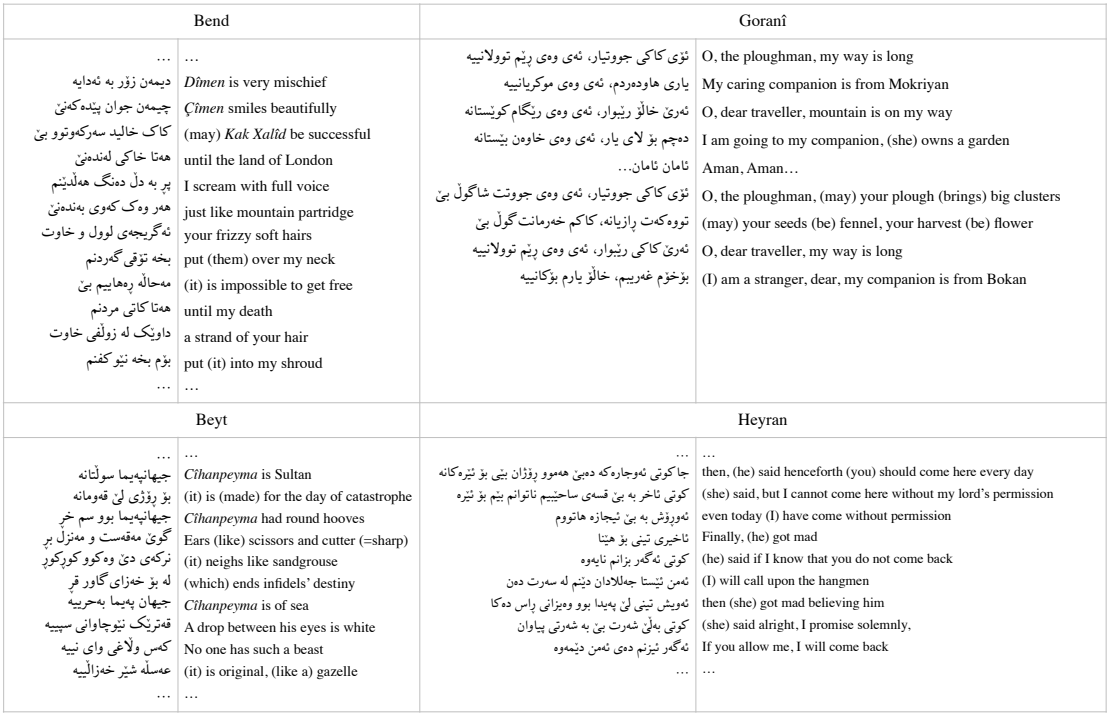
2.3. Goranî

In addition to a specific genre, the term *Goranî* is also one of the words for "song" in Sorani Kurdish. It should not be confused with the Goranî dialect[1]. There is a fine line between what is referred to with the term Goranî and with other terms such as *Stran*, *Beste*, and *Meqam* (Salimi, 2015) in different dialects and regions. The terms are observed to be used interchangeably across the Kurdish speaking areas regardless of the dominant dialect.
The themes of *Goranî* come from diverse contexts. This diversity creates different types of *Goranî* for various occasions such as wedding, birth, feasts and funeral, and various feelings, such as love, happiness and hope (Broughton et al., 2006).
A special form of *Goranî* is *Meqam* which has different characteristics among the speakers of different Kurdish dialects. For example, it is essentially used in religious practices in some Kurdish groups (Merati, 2015), while it is a special lyric whose main motif is a love story among other groups. It is usually performed without musical accompaniment.

2.4. Heyran

It is a form of lyric which mostly tells love stories, but it could also be about tragic stories and actions of heroes in the battles (Merati, 2015). The *Heyranbêj*, literally meaning the sayer of Heyran, is the one who performs Heyran. According to Merati (2015), it is a lyric form which is performed in Sorani and mostly in the Iraqi and Iranian Kurdistan. In the Kurdistan Region of Iraq, this type of lyric is

[1] also written as Guranî.



Bend		Goranî	
…	…	ئۆی کاکی جووتیار، ئەی وەی ڕێم تووڵانییە	O, the ploughman, my way is long
دیمەن زۆر بە ئەدایە	*Dîmen* is very mischief	یاری هاودەردم، ئەی وەی موکریانییە	My caring companion is from Mokriyan
چیمەن جوان پێدەکەنێ	*Çîmen* smiles beautifully	ئەرێ خاڵۆ ڕێبوار، ئەی وەی ڕێگام کوێستانە	O, dear traveller, mountain is on my way
کاک خالید سەرکەوتوو بێ	(may) *Kak Xalîd* be successful	دەچم بۆ لای یار، ئەی وەی خاوەن بێستانە	I am going to my companion, (she) owns a garden
هەتا خاکی لەندەنێ	until the land of London	ئامان ئامان…	Aman, Aman…
پڕ بە دڵ دەنگ هەڵدێنم	I scream with full voice	ئۆی کاکی جووتیار، ئەی وەی جووتت شاگوڵ بێ	O, the ploughman, (may) your plough (brings) big clusters
هەر وەک کەوی بەندەنێ	just like mountain partridge	تووەکەت ڕازیانە، کاکم خەرمانت گوڵ بێ	(may) your seeds (be) fennel, your harvest (be) flower
ئەگریجەی لوول و خاوت	your frizzy soft hairs	ئەرێ کاکی ڕێبوار، ئەی وەی ڕێم تووڵانییە	O, dear traveller, my way is long
بخە تۆقی گەردنم	put (them) over my neck	بۆخۆم غەریبم، خاڵۆ یارم بۆکانییە	(I) am a stranger, dear, my companion is from Bokan
مەحاڵە ڕەهاییم بێ	(it) is impossible to get free		
هەتا کاتی مردنم	until my death		
داوێک لە زوڵفی خاوت	a strand of your hair		
بۆم بخە نێو کفنم	put (it) into my shroud		
…	…		

Beyt		Heyran	
…	…	…	…
جیهانپەیما سوڵتانە	*Cîhanpeyma* is Sultan	جا کوتی ئەوجارەکە دەبێ هەموو ڕۆژان بێی بۆ ئێرەکانە	then, (he) said henceforth (you) should come here every day
بۆ ڕۆژی لێ قەومانە	(it) is (made) for the day of catastrophe	کوتی ئاخر بە بێ قسەی ساحێبیم ناتوانم بێم بۆ ئێرە	(she) said, but I cannot come here without my lord's permission
جیهانپەیما بوو سم خڕ	*Cîhanpeyma* had round hooves	ئەورۆش بە بێ ئیجازە هاتووم	even today (I) have come without permission
گوێ مەقەست و مەنزڵ بڕ	Ears (like) scissors and cutter (=sharp)	ئاخیری تێنی بۆ هێنا	Finally, (he) got mad
نرکەی دێ وەکوو کوڕکوڕ	(it) neighs like sandgrouse	کوتی ئەگەر بزانم نایەوە	(he) said if I know that you do not come back
لە بۆ خەزای گاور قڕ	(which) ends infidels' destiny	ئەمن ئێستا جەللادان دێنم لە سەرت دەن	(I) will call upon the hangmen
جیهان پەیما بەحرییە	*Cîhanpeyma* is of sea	ئەویش تێنی لێ پەیدا بوو وەیزانی ڕاس دەکا	then (she) got mad believing him
قەترێک نێوچاوانی سپییە	A drop between his eyes is white	کوتی بەڵێ شەرت بێ بە شەرتی پیاوان	(she) said alright, I promise solemnly,
کەس وڵاغی وای نییە	No one has such a beast	ئەگەر ئیزنم دەی ئەمن دێمەوە	If you allow me, I will come back
عەسڵە شێر خەزاڵییە	(it) is original, (like a) gazelle	…	…
…	…		

Figure 1: A comparison of the four genres included in our corpus. The translations are literal and additional words are provided in parentheses. Proper names are italicized.

also called by the same name. However, a similar type is called *Lawik* in the Kurmanji-speaking areas of Iraq which is usually longer than Heyran (Merati, 2015). According to Merati (2015), the stanza of Heyran is constructed on three verses with three rhymes in each verse.

2.5. Other Forms of Kurdish Lyrics

As it was mentioned earlier, Kurdish lyrics are not restricted to the forms presented here. They are diverse in their form, colorful in their themes, and varied in their subjects and contexts. Some of these forms are particular to certain dialects, while some are common among the dialects.

One example of these forms is *Hore*. It is particular to Hawramî (Goranî) and the Southern Kurdish dialects, spoken in the Kurdish speaking regions of Iran and Iraq. *Hore* is assumed to be a type of singing with more than several thousand years of history (Merati, 2015). Another example is *Çamary*, which is a song in mourning circumstances, particularly, for the death of socially important individuals (Merati, 2015).

3. Related Work

In this section, we address the related work regarding the collection of folkloric content and lyrics as resources in other Kurdish dialects and also other languages.

Regarding Kurdish, Hamelink and Barış (2014) created a corpus from Kurmanji lyrics. This corpus includes 84 *Kilam*s (or *Kelams*) (Merati, 2015) which is a title for a type of music mostly in Kurmanji speaking areas, though with different attributes, depending on the geographical position of the community in which the music is performed (Hamelink and Barış, 2014).

Regarding other languages, a famous work on the English song lyrics is The Million Song dataset (Mahieux et al., 2011) which is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. Moreover, Taft (1977) collected over two thousand texts which were performed by about 350 blues (an African-American music genre) singers. Mahedero et al. (2005) presents experiments on lyrics using NLP methods to identify lyrics' language, thematic categorization, structure extraction and to perform similarity searches. This research suggested that information which acoustic and cultural metadata would have been providing could be further improved when they are accompanied by lyrics. McNeil (2018) reports the collection of folklore poetry and popular song lyrics alongside other forms to develop a Tunisian Arabic corpus.

The application of lyric processing in music analysis have been investigated from a variety of perspectives. For example, Hu et al. (2009) conducted research to examine the role that lyric texts could play in the mood classification in audio music. They found the lyric features can outperform audio features in the classification of mood categories in certain cases. They also found that combining lyrics and audio features improve performances on a majority of mood classification categories.

Also, Rodrigues et al. (2019) developed a corpus of English lyrics which they expected that would assist in testing and

evaluation of tools pertinent to the language generation in the poetry and lyrics context.

In the same vein, the International Workshop on Folk Music Analysis (FMA)[2] is an annual workshop dedicated to folk music analysis since 2011. The computational folklorisitics has been a repetitive theme in this workshop series wherein many scholars have reported on variety of corpora which have been developed based on various ethnic and national folklore (for example, see (Holzapfel, 2014), (Beauguitte et al., 2016), and (Ali-MacLachlan and Hockman, 2019)). Although the dominant area of these workshops is about musicology, some attempts concerning language processing are observed. For instance, (Strle and Marolt, 2014) reported on the collection of 1,965 variants of Slovenian folk narrative poems to evaluate the effectiveness of two different methods of semantic analysis in NLP.

4. Lyrics Corpus

We transcribed a set of 162 songs in various genres in the four types of Kurdish folkloric materials: *Bend, Beyt, Goranî*, and *Heyran*. Given the wide range of Kurdish dialects and sub-dialects, we only focused on the Sorani dialect of Kurdish which is mostly spoken in the Kurdish regions in Iran and Iraq. As a song may have been performed by many singers, we considered the recording quality and authenticity of the lyrics as the criteria to select one.

4.1. Text Transcription

The transcription process was carried out by native Kurdish speakers by listening to the audiovisual materials. In order to find such content easily and also receive feedback from others regarding our transcription quality, we created a channel on the Telegram Messenger[3] where we regularly published the lyrics along with the audiovisual material over four months. Table 1 provides the statistics of the corpus.

Two main challenges in transcribing the lyrics were the quality of the recording, which was low in some cases, and the way that the singer articulated the words. In many cases, we observed that some words were not pronounced clearly and some parts of the lyrics are left incomplete due to the rhythm. In such cases, we tried to write the lyrics based on various performances of the same song.

Genre	Number of songs	Number of tokens (characters)
Bend	12	6455 (56,723)
Beyt	8	17994 (200,981)
Goranî	141	22588 (212,408)
Heyran	1	2545 (2273)
Total	162	49,582 (472,385)

Table 1: Statistics of the lyrics corpus

4.2. Conversion to TEI

We converted the transcribed songs into TEI format, which is based on XML. The XML format provides a structured form to represent segments in the lyrics and metadata of each song, including song name, singer's name, URL to the audio file, song ID, and the type of the song. Regarding the name of the songs, we used a title that is most frequently known to the public. We used the refrains to give the title to the songs for which we could not find any title. However, some of the Bends left without a title due to lack of a refrain or a popular title. Figure 2 presents the XML structure of a song of Goranî genre. It should be noted that the attributes are customized and are not defined elsewhere in TEI.

Some of the lyrics are composed of classical Kurdish poems. We use `type="poem"` attribute to distinguish these parts from the folkloric lyrics. In addition, Beyt and Heyran performers usually provide comments in plain language to facilitate the comprehension of the story and guarantee the story flow. We use `type="comment"` to highlight performer's comments.

```
<text id="131">
  <div type="song">
      <head>هەرزالێ</head>
      <singer>ناسر ڕەزازی</singer>
      <audio>Audio/Gorani/131.mp3</audio>
      <lg type="Gorani">
          <l>هەرزالێ، براکەم هەرزالێ</l>
          <l>هەرزاڵ چاترە لە مالێ</l>
          <l>ئەو چۆپیکێشە چۆپی ئەکێشێ</l>
          <l>لە خوام گەرەکە، هیچکوێنەی نەیشێ</l>
          <l>هەرزالێ، براکەم هەرزالێ</l>
          <l>هەرزاڵ چاترە لە مالێ</l>
          <l>یان بە گوڵباخی یان بە گوڵەزەردە</l>
          <l>یان بە نازی خۆت بمکە پەروەردە</l>
          <l>هەرزالێ، براکەم هەرزالێ</l>
          <l>هەرزاڵ چاترە لە مالێ</l>
          <l>توخوا چۆپیکێش چۆپیت بە لەنگەر</l>
          <l>هەر وەک پێشمەرگە ڕوو بکە لە سەنگەر</l>
      </lg>
  </div>
```

Figure 2: A transcribed song converted in TEI

5. Evaluation

In addition to the statistics of the corpus in Table 1, we evaluate the content by comparing it with two other Sorani Kurdish corpora, Pewan (Esmaili and Salavati, 2013) and KTC (Abdulrahman et al., 2019) which are respectively general-purpose and domain-specific.

Calculating the frequency of words is a measure to understand how they semantically form the resources. Table 2 presents the ten most frequent tokens in our corpus and the two other Sorani Kurdish corpora. Although all these words are function words, i.e., a word whose purpose is to contribute to the syntax rather than the meaning of a sentence, they are not similarly distributed in the lyrics against the two other resources. The frequency of pronouns is observed in the lyrics text, which indicates the narrative nature of the folkloric songs. In addition, punctuation signs, which are commonly used in formal writing in the two other resources, are not frequently used in the lyrics.

[2] `http://www.folkmusicanalysis.org/`

[3] Available at `https://t.me/Folklorelyrics`

In the same vein, Table 3 provides the ten most frequent words excluding the function words. The Pewan corpus has words associated with politics, as it was created based on the news articles. On the other hand, KTC has a more diverse range of words since it contains many domain-specific topics, from geography to linguistics and theology. Regarding the lyrics corpus, the most frequent non-function words are oriented around poetic and literary themes. Moreover, lyrics vocabulary can be used to analyze the semantic change thanks to archaism.

One other evaluation measure is linguistic representativeness (Gray et al., 2017). As the lyrics corpus contains various Sorani sub-dialects, various dialectal differences in the lexical choice and morphology are observed. Among the non-function words, we counted 7,316 tokens in the lyrics which do not exist among the 946,569 unique tokens of a basic Sorani Kurdish dictionary (Ahmadi et al., 2019) and the two other corpora. Having said that, considering lemmatization, which was not possible due to lack of tools for Kurdish, we expect that this number of words decreases to some extent, but still leaving a considerable number of words that could be added to the dictionaries.

Our corpus	KTC (Abdulrahman et al., 2019)	Pewan (Esmaili et al., 2013)
لە (from)	**و**	**لە**
بە (to)	**لە**	**و**
و (and)	**بە**	**بە**
بۆ (for)	که (that)	**بۆ**
بێ (without)	**بۆ**	/
ئەو (she/he/it/that)	**ئەو**	که (that)
من (I/me)	ئەم (this/it)	**ئەو**
تۆ (you)	.	-
ئەی/ۆەی/ئای (O, oh)	:	:
هەر (only, each)	،	]

Table 2: The 10 most frequent tokens in our corpus versus two other Sorani Kurdish corpora. The common tokens are highlighted in bold.

Our corpus	KTC (Abdulrahman et al., 2019)	Pewan (Esmaili et al., 2013)
گیان (soul, dear (adjective))	مرۆڤ (human)	کوردستان (Kurdistan)
دەڵێم ((I) say)	گەورە (big)	عێراق (Iraq)
دەکەم ((I) do)	کورد (Kurd, Kurdish)	هەرێمی (region of)
دەبێ ((it) should)	خودای (god, god of)	سەرۆکی (president of)
وەرە (come (imperative))	کوردستان (Kurdistan)	هەولێر (Erbil)
کرد ((she/he/it) did)	واتە ((it) means)	حکومەتی (government of)
دڵ (heart)	زمانی (language of)	شاری (city of)
گوڵ (flower)	پێویستە ((it) is needed)	ئێران (Iran)
شەو (night)	مافی (right of)	ئەنجومەنی (parliament of)
خۆم (myself)	وزەی (energy of)	ئەمەریکا (USA)

Table 3: The 10 most frequent tokens, excluding function words, in our corpus versus two other Sorani Kurdish corpora

6. Conclusion

We presented a corpus of folkloric lyrics in the Sorani dialect of Kurdish. The corpus contains lyrics of 162 songs (49,582 tokens) in four Kurdish musical genres: 12, 141, 8 and 1 songs in Bend, Goranî, Beyt and Heyran, respectively. We demonstrated that the current resource provides additional linguistic information, which is not represented in other Sorani Kurdish corpora.

This work is initial in using Sorani lyrics as a source for Kurdish language processing. Therefore, numerous areas could be counted for further developments, such as named-entity recognition, relation extraction, computational musicology and co-reference resolution. Another future work could be the enrichment of this corpus by adding content in other Kurdish dialects and by translating them into other languages, particularly English. We believe that development of such resources will pave the way for further developments in Kurdish language processing, therefore helping it to become a resourceful language. Since the Arabic script of Kurdish has proved to pose challenges in Kurdish text processing (Ahmadi, 2019), we would suggest the transliteration of the corpus into the Latin script, as it is also mostly used in the Kurmanji dialect.

The corpus is publicly available for non-commercial use under the CC BY-NC-SA 4.0 license at `https://github.com/KurdishBLARK/KurdishLyricsCorpus` [4].

7. Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful suggestions and careful reading of the manuscript.

8. Bibliographical References

Abdulrahman, R., Hassani, H., and Ahmadi, S. (2019). Developing a fine-grained corpus for a less-resourced language: the case of Kurdish. *WiNLP ACL 2019*.

Abello, J., Broadwell, P., and Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM*, 55(7):60–70.

Abubakir, N. H. R. (2016). *Bringing Kurdish Music to the West*. Ph.D. thesis, University of Kansas.

Ahmadi, S., Hassani, H., and McCrae, J. P. (2019). Towards Electronic Lexicography for the Kurdish Language. In *Proceedings of the eLex 2019 conference*, pages 881–906, Sintra, Portugal, 1–3 October. Brno: Lexical Computing CZ, s.r.o.

Ahmadi, S. (2019). A rule-based Kurdish text transliteration system. *Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):18:1–18:8.

Ali-MacLachlan, I. and Hockman, J. (2019). *Proceedings of the 9th International Workshop on Folk Music Analysis (FMA2019), 2-4 July, 2019*. Birmingham City University.

Allison, C. (2001). *The Yezidi Oral Tradition in Iraqi Kurdistan*. Routledge.

Ataman, D. (2018). Bianet: A Parallel News Corpus in Turkish, Kurdish and English. *arXiv preprint arXiv:1805.05095*.

[4] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

Barzegar Khaleghi, M. (2009). Kurdish Myths and Legends. *Kavoshnameh - Journal of Research in Persian Language and Literature*, 18:201–223. [In Farsi].

Beauguitte, P., Duggan, B., and Kelleher, J. (2016). *Proceedings of the 6th International Workshop on Folk Music Analysis, 15-17 June, 2016*. Dublin Institute of Technology.

Blum, S. and Hassanpour, A. (1996). 'the morning of freedom rose up': Kurdish popular song and the exigencies of cultural survival. *Popular Music*, 15(3):325–343.

Bocheńska, J. (2014). Kurdish Contemporary Literature in Search for Ordo Amoris-Some Reflections on the Continuity of the Kurdish Literary Tradition and Ethics. *Nûbihar Akademî*, 1(1):35–54.

Brenneman, R. L. (2016). *As strong as the mountains: A Kurdish cultural journey*. Waveland Press.

Broughton, S., Ellingham, M., Lusk, J., and Clark, D. A. (2006). *The Rough Guide to World Music: Africa & Middle East*, volume 1. Rough Guides.

Christensen, D. (2007). Music in Kurdish identity formations. In *Conference on Music in the World of Islam. Assilah*, pages 8–13.

Esmaili, K. S. and Salavati, S. (2013). Sorani Kurdish versus Kurmanji Kurdish: An Empirical Comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 300–305.

Esmaili, K. S., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S., and Hakimi, S. (2013). Building a test collection for Sorani Kurdish. In *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*, pages 1–7. IEEE.

Gray, B., Egbert, J., and Biber, D. (2017). Exploring methods for evaluating corpus representativeness. *the Corpus Linguistics International Conference 2017. Birmingham, UK*.

Hamelink, W. and Barış, H. (2014). Dengbêjs on borderlands: Borders and the state as seen through the eyes of Kurdish singer-poets. *Kurdish Studies*, 2(1):34–60.

Hamelink, W. (2016). *The Sung Home. Narrative, Morality, and the Kurdish Nation*. Brill.

Hassani, H. (2018). Blark for multi-dialect languages: towards the Kurdish BLARK. *Language Resources and Evaluation*, 52(2):625–644.

Hassanpour, A. (2005). Wanderings in Adalar Sahilinde. In *Joyce Blau l'éternelle chez les Kurdes*, pages 62–73. Institut français d'études anatoliennes. [Online; accessed 09-Nov-2019].

Holzapfel, A. (2014). *Proceedings of the Fourth International Workshop on Folk Music Analysis, 12 and 13 June, 2014, Istanbul, Turkey: FMA2014*. Boğaziçi University.

Hu, X., Downie, J. S., and Ehmann, A. F. (2009). Lyric text mining in music mood classification. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, pages 411–416.

Ide, N. and Véronis, J. (1995). *Text encoding initiative: Background and contexts*, volume 29. Springer Science & Business Media.

Kreyenbroek, P. G. (2005). Kurdish written literature. *Encyclopædia Iranica*, page 2.

Leezenberg, M. et al. (2011). Soviet Kurdology and Kurdish Orientalism. *2011). The Heritage of Soviet Oriental Studies*, pages 86–102.

Mahedero, J. P., MartÍnez, Á., Cano, P., Koppenberger, M., and Gouyon, F. (2005). Natural language processing of lyrics. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 475–478. ACM.

Mahieux, T. B., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. *ISMIR-11*.

Malmasi, S. (2016). Subdialectal differences in Sorani Kurdish. In *Proceedings of the third workshop on nlp for similar languages, varieties and dialects (vardial3)*, pages 89–96.

McNeil, K. (2018). Tunisian arabic corpus: Creating a written corpus of an 'unwritten'language. *Arabic Corpus Linguistics*, page 30.

Merati, M. A. (2015). *Les formes fondamentales de la musique kurde d'Iran et d'Irak : hore, siâ-çamane, danses, maqâm [In English: The basic forms of Kurdish music from Iran and Iraq]*. Ph.D. thesis, Paris Nanterre University.

Mikailee, H. (2015). "Beyt" in Kurdish Folk Literature. *Journal of Kurdish Literature*, 1(1):57–82. [In Farsi].

Rasul, E. M. (1999). *A Research in the Kurdish Folklore*. Salhaddin Ayyobi. [Farsi translation].

Reigle, R. F. (2014). A brief history of Kurdish music recordings in Turkey. *Hellenic Journal of Music, Education and Culture*, 4(1).

Rodrigues, M. A. G., de Paiva Oliveira, A., and Moreira, A. (2019). Development of a song lyric corpus for the english language. In *International Conference on Applications of Natural Language to Information Systems*, pages 376–383. Springer.

Salimi, H. (2015). Taking a look at Kurdish Folklore. *National Studies Quarterly*, 8(2). [In Farsi].

Sharifi, A. (2005). Kurdish Myths and Legends. *Iranian People's Culture Quarterly*, 7–8. [In Farsi].

Strle, G. and Marolt, M. (2014). Uncovering semantic structures within folk song lyrics. In *Workshop on Folk Music Analysis (FMA2014)*, page 40.

Taft, M. (1977). *The lyrics of race record blues, 1920-1942: a semantic approach to the structural analysis of a formulaic system*. Ph.D. thesis, Memorial University of Newfoundland.

9. Language Resource References

Abdulrahman, Roshna and Hassani, Hossein and Ahmadi, Sina. (2019). *Developing a Fine-grained Corpus for a Less-resourced Language: the case of Kurdish*. Kurdish BLARK.