Towards Electronic Lexicography for the Kurdish Language

Sina Ahmadi[1], Hossein Hassani[2], John P. McCrae[1]

[1] Insight Centre for Data Analytics, National University of Ireland Galway
[2] Department of Computer Science and Engineering, University of Kurdistan Hewlêr
E-mail: sina.ahmadi, john.mccrae@insight-centre.org, hosseinh@ukh.edu.krd

Abstract

This paper describes the development of lexicographic resources for Kurdish and provides a lexical model for this language. Kurdish is considered a less-resourced language, and currently, lacks machine-readable lexical resources. The unique potential which Linked Data and the Semantic Web offer to e-lexicography enables interoperability across lexical resources by elevating the traditional linguistic data to machine-processable semantic formats. Therefore, we present our lexicon in Ontolex-Lemon ontology as a standard model for sharing lexical information on the Semantic Web. The research covers the Sorani, Kurmanji, and Hawrami dialects of Kurdish. This research suggests that although Kurdish is a less-resourced language, in terms of documented lexicons, it has a wide range of resources, but because they are not machine-readable they could not contribute to the language processing. The outcome of this project, which is made publicly available, assists scholars in their efforts towards making Kurdish a resource-rich language.

Keywords: Kurdish; e-lexicography; less-resourced languages; machine-readable dictionary

1. Introduction

Linguistic resources are knowledge repositories which not only provide lexical and semantic descriptions of words but also reflect the culture and civilization of speakers of a language. In an era when human language is more and more frequently processed by machines, such resources are crucial components of language technology and natural language processing (NLP). Kurdish, as a less-resourced Indo-European language spoken in several dialects and written using different scripts (Forcada et al., 2019), still lacks such resources. In an attempt to remedy the lack of resources for Kurdish, we provide machine-readable dictionaries for three of the five main dialects of Kurdish, namely Kurmanji, Sorani, and Hawrami.

A machine-readable dictionary (MRD) not only provides lexicographic information in an electronic form, but is also a database which can be queried and therefore integrated in NLP tools. As the body of the research in Kurdish language processing is still scant, we believe that such resources will pave the way for further developments in the field. We also believe that lexical resources will enable researchers to address more NLP tasks which may require lexicographic resources such as word sense disambiguation (Navigli & Ponzetto, 2012) and semantic parsing (Shi & Mihalcea, 2005) and enhance the quality of the existing NLP applications.

The Semantic Web as an extension of the World Wide Web (WWW) represents an effective means of data representation and enables users and computers to retrieve and share information efficiently (Berners-Lee et al., 2001). The Resource Description Framework (RDF) is the foundational data model for the Semantic Web. Unlike traditional databases where data has to adhere to a fixed schema, RDF documents are not prescribed by a schema and can be described without additional information, making RDF data model self-describing (Klyne & Carroll, 2004). More recently, the concept of the Web of Linked Data, which makes RDF data available using the HyperText Transfer Protocol (HTTP) and Linguistic Linked Open Data (Chiarcos et al., 2013), has gained traction along with the Semantic Web, particularly in the NLP community as a standard for linguistic resource creation. Moreover, the unique potential which the Semantic Web and Linked Data offer to e-lexicography enables interoperability across lexical resources by leveraging printed or unstructured linguistic data to machine-readable semantic formats.

This paper has two major contributions:

- It provides a thorough review of the current state of Kurdish lexicography, both traditional and electronic. Such a review includes an analysis of the properties of the existing Kurdish dictionaries, such as type of dictionary (monolingual, bilingual, multilingual), script of the Kurdish text (Persian-Arabic, Latin or Cyrillic), description of the content and size of dictionaries. Although very few in comparison to printed dictionaries, terminological resources and electronic dictionaries are also covered in this paper. This review helped us to differentiate between the lack of resources and unavailability of lexicographic resources in electronic forms. We discovered that Kurdish, from the lexicographic point of view, is not as less-resourced as claimed in the literature. Instead, other issues have hindered the availability of these resources in a machine-readable form, which has resulted in the perception that the language lacks such essential assets. This is not equally true for all the Kurdish dialects, but it is obvious for the two widely spoken dialects, namely Kurmanji and Sorani.
- We present three machine-readable dictionaries based on the OntoLex-Lemon model for Kurmanji, Sorani and Hawrami. We not only included frequent headwords in the dictionaries, namely 4,172 entries for Kurmanji, 5,683 entries for Sorani and 1,184 for Hawrami, but also tried to create a prototypical resource which may be easily adapted by future Kurdish lexicographers. Despite the existence of a few electronic word lists and glossaries for Sorani and Kurmanji, our electronic Hawrami dictionary is the first one of its kind for this dialect.

For this, we consider two stages in the development of our resources. First, we collect the vocabulary for each dialect. This stage includes manual work for the extraction of entries and annotating each part of their description, such as gender, part-of-speech (PoS), sense, English translations, example and etymology. This step is followed by a semiautomatic normalization of the scripts and orthography. In the next step, the

lexicographic information is semi-automatically transformed from a tabular format into the OntoLex-Lemon model in the Resource Description Framework (RDF).

The rest of this paper is organized as follows. We first describe the Kurdish language, its various dialects and scripts in Section 2. In Section 3, we provide a survey on the history of Kurdish lexicography and available lexicons. Section 4 describes the development of our resources according to the OntoLex-Lemon standard. Following this section, insights into the developed resources are provided in Section 5. The paper is concluded in Section 6, where we provide suggestions for modern e-lexicography for Kurdish and future steps in this direction. Note that throughout this study, lexicon and dictionary are used interchangeably.

2. Kurdish language

2.1 Dialects

Kurdish is an Indo-European multi-dialect language which is spoken by about 30 million speakers (Hassani, 2018). The dialects are referred by different names, namely Kurmanji, Sorani, Hawrami and Kirmashani (Hassani, 2018). The Kurmanji speakers, as the majority of Kurdish speaking population, are located in different areas of Syria, Iraq, Turkey and Iran. Sorani is the second most popular dialect, which is mainly spoken among Kurds in Iran and Iraq. Similarly, Hawrami is primarily spoken in Iran and Iraq, but among a smaller community. Moreover, almost all Kurdish dialects are also spoken among a large Kurdish diaspora in different western countries (Hassanpour, 1992).

The debate over the concept of dialects versus languages, the attribution of different dialects to Kurdish or considering some as separate languages has been around for decades (Hassani, 2018). According to the literature (Hassanpour, 1992; Haig & Matras, 2002; Hassani, 2018), the debate expanded to how to categorize and name the dialects. However, to avoid drifting beyond our purpose, in this research we prefer to follow the common approach among the researchers in Kurdish NLP with regard to dialect attribution, their categorization, and naming style according to the way presented in the Kurdish BLARK (Hassani, 2018).

2.2 Scripts and orthographies

Kurdish poetry and prose narratives were historically transmitted orally (Kreyenbroek, 2005), therefore the language does not have a long history of written texts (Hassani & Medjedovic, 2016). While some scholars have different opinions, the dominant conclusion dates the appearance of the first written Kurdish text to circa 1600 (Hassani, 2018). Since then, the language has been written in Persian-Arabic until the beginning of the 20th century, when due to geopolitical conditions the usage of Latin, Cyrillic, and to a limited extent, Armenian scripts was started. In the 1920s, the first attempts to present a standard writing system for Kurdish began. As a result, in 1932 Jeladet Ali Bedirkhan (in Kurdish, *Celadet Elî Bedirxan*) introduced a

Latin-based orthography (also known as Bedirxan alphabet) (Bedirxan & Lescot, 1970), while a group of scholars introduced one based on the Persian-Arabic script in Iraq. These orthographies are both based on the phonetics of the language. The usage of Cyrillic and Armenian was mainly restricted to the communities in Armenia and the former Soviet countries (Hassanpour, 1992). Gradually, the Persian-Arabic and Latin-based scripts have become more dominant in various Kurdish speaking regions, although their popularity differs from region to region. The Persian-Arabic orthography is dominant in the Kurdish regions of Iraq, Iran, and Syria (Haig & Matras, 2002). On the other hand, the Latin-based orthography is used by the Kurds in Turkey. According to Hassani and Medjedovic (2016), the usage of Latin-based orthography is growing and becoming more popular in Iraq and Syria, with a greater usage by the Kurdish media, particularly in the Kurdistan Region of Iraq.

In an attempt to standardize and unify the scripts for all Kurdish dialects, the Kurdish Academy of Language has recently introduced a Unified Kurdish Alphabet, *Yekgirtû* [1], which is based on the Latin orthography. Figure 1 illustrates Kurdish phonemes in all dialects and their corresponding letters in the alphabets. The grey cases refer to non-existing characters.

2.3 Kurdish language processing

Hassani (2018) provides a summary of the Kurdish NLP situation in which the status of the available data and tools for Kurdish NLP are presented. However, we also address a few essential efforts on Kurdish NLP which are pertinent to the current research, particularly on Kurdish language processing resources and tools.

The initiative to create corpus for Kurdish dates back to 1998 (Gautier, 1998). However, efforts in creating machine-readable corpora for Kurdish are recent. The first machine-readable corpus for Kurdish is the Leipzig Corpora Collection which contains some 56,000 sentences of Sorani Kurdish constructed using different sources such as the Internet, newspapers, and Wikipedia (Biemann et al., 2007). In 2013, the Kurdish Language Processing Project created Pewan (Esmaili et al., 2013) which is composed of 115,000 Sorani and 25,000 Kurmanji news articles. KurdNet (Aliabadi et al., 2014) is the Kurdish WordNet, and currently only contains Sorani translations of the Base Concept of the English WordNet (Miller, 1995). Bianet is a parallel news corpus of Turkish, English and Kurmanji containing 3,214 articles (Ataman, 2018). In addition, researchers have created Kurdish corpora for particular NLP tasks, for example, part-of-speech (PoS) annotation (Walther & Sagot, 2010; Walther et al., 2010), dialectology (Hassani & Medjedovic, 2016; Malmasi, 2016), creating dependency treebanks (Gökırmak & Tyers, 2017), and intralanguage and interlanguage machine translation (Hassani, 2018; Ahmadi, 2019).

[1] http://kurdishacademy.org/?p=111

Kurdish phonemes (IPA)	Our suggestion	Latin-based	Yekgirtú	Persian-Arabic-based			
				Initial	Middle	Final	Single
[aː]	A a	A a	A a	ئا	ـا	ـا	ا
[b]	B b	B b	B b	بـ	ـبـ	ـب	ب
[t͡ʃ]	Ç ç	Ç ç	C c	چـ	ـچـ	ـچ	چ
[d͡ʒ]	C c	C c	J j	جـ	ـجـ	ـج	ج
[d]	D d	D d	D d	دـ	ـدـ	ـد	د
[æ]	E e	E e	E e	ئەـ	ـەـ	ـە	ە
[eː]	Ê ê	Ê ê	É é	ئێـ	ـێـ	ـێ	ێ
[f]	F f	F f	F f	فـ	ـفـ	ـف	ف
[g]	G g	G g	G g	گـ	ـگـ	ـگ	گ
[h]	H h	H h	H h	هـ	ـهـ	ـه	ه
[ɪ]	I i	I i	I i				
[iː]	Î î	Î î	Í í	ئیـ	ـیـ	ـی	ی
[ʒ]	J j	J j	Jh jh	ژـ	ـژـ	ـژ	ژ
[k]	K k	K k	K k	کـ	ـکـ	ـک	ک
[l]	L l	L l	L l	لـ	ـلـ	ـل	ل
[ɬ]	Ł ł	Ll ll	Ll ll	ڵـ	ـڵـ	ـڵ	ڵ
[m]	M m	M m	M m	مـ	ـمـ	ـم	م
[n]	N n	N n	N n	نـ	ـنـ	ـن	ن
[oː]	O o	O o	O o	ئۆ	ـۆـ	ـۆ	ۆ
[p]	P p	P p	P p	پـ	ـپـ	ـپ	پ
[q]	Q q	Q q	Q q	قـ	ـقـ	ـق	ق
[ɾ]	R r	R r	R r	ڕـ	ـڕـ	ـڕ	ڕ
[r]	Ř ř	Rr rr	Rr rr	ڕـ	ـڕـ	ـڕ	ڕ
[s]	S s	S s	S s	سـ	ـسـ	ـس	س
[ʃ]	Ş ş	Ş ş	Sh sh	شـ	ـشـ	ـش	ش
[t]	T t	T t	T t	تـ	ـتـ	ـت	ت
[ʊ]	U u	U u	U u	ئو	ـو	ـو	و
[uː]	Û û	Û û	Ú ú	ئوو	ـووـ	ـوو	وو
[v]	V v	V v	V v	ڤـ	ـڤـ	ـڤ	ڤ
[w]	W w	W w	W w	وـ	ـوـ	ـو	و
[x]	X x	X x	X x	خـ	ـخـ	ـخ	خ
[j]	Y y	Y y	Y y	یـ	ـیـ	ـی	ی
[z]	Z z	Z z	Z z	زـ	ـزـ	ـز	ز
[ħ]	Ḧ ḧ		H', h'	حـ	ـحـ	ـح	ح
[ʕ]	Ë ë		'	عـ	ـعـ	ـع	ع
[ɣ]	Ẍ ẍ		X', x'	غـ	ـغـ	ـغ	غ
[ʉː]	Ü ü		Ù ù	ئۆ	ـۆـ	ـۆ	ۆ
[ɣ]	D̂ d̂			ڎـ	ـڎـ	ـڎ	ڎ
[ʔ]	'			ئـ	ـئـ	ـئ	ئ
[ʁ]	Ğ ğ						

Figure 1: A comparison of the alphabets used for Kurdish writing.
A unified script for all dialects is suggested.

Kurdish NLP is a young sector in the realm of worldwide NLP. Particularly, to be able to prepare the underlying resources to leverage its language processing capacity, it needs a wide range of tools such as Optical Character Recognition (OCR), thesauri, treebanks, machine-readable lexicons, a variety of language models, and transliterators

for its various scripts, to name a few. However, currently, most of these tools either do not exist or they are in their infancy. The situation and the requirements have been addressed by several researchers (Hassani, 2018; Yaseen & Hassani, 2018; Ahmadi, 2019). The current research is an attempt to improve this situation.

3. Kurdish Lexicography

Since poems have historically had a special place in Kurdish literature, the earliest works in Kurdish lexical studies were in verse. *Nûbihara Biçûkan* (*The Kids' Spring*) which dates back to 1683, is considered the first Kurdish dictionary and the first Kurdish work in children's literature (Yıldırım, 2008). This resource contains 1,000 Kurdish-Arabic pairs which were taught for years at Kurdish elementary schools to teach Arabic for Koranic studies (Hassanpour, 1992). Poetic resources have been historically used among the Kurds for educational purposes as the translations are provided in rhythm. Bolelli and Ertekin (2017) count eight poetic resources for various Kurdish dialects, which mostly provide Arabic translations. Recently, Ertekin (2017) presented a Turkish-Kurmanji dictionary in verse. The following is an example from the Nodeyî (1936) which was created according to Yıldırım (2008) in verse in Sorani Kurdish:

(أین) لەکوێ (مَن) کێیە (أَیَّانَ) کەی	(أین) lekwê (مَن) kê ye (أَیَّانَ) key	(أین) where (مَن) is who (أَیَّانَ) when
(إن) ئەگەر (ما) چی (مَتَی) کەی (سَمّ) ژار	(إن) eger (ما) çî (مَتَی) key (سَمّ) jar	(إن) if (ما) what (مَتَی) when (سَمّ) poison

Figure 2: A couplet from (Nodeyî, 1936) Arabic-Sorani Kurdish work (original on the left, transliterated to Latin in the middle, translated into English on the right). Kurdish words appear in parentheses immediately after the source words in Arabic.

Despite the historical popularity of poetic resources in traditional Kurdish schools, such resources can hardly be categorized as dictionaries due to the superficial representation of lexical information and the poetic structure. Moreover, these resources cannot be consulted, and therefore it is impossible to systematically retrieve data from them.

In this section, we describe some of the existing lexicographic resources for Kurdish which have played an essential role in forming Kurdish lexicography. A complete list of Kurdish lexicographic resources is provided in Appendix A. We have not considered word lists and glossaries which appear as part of other works in linguistics and literature (e.g. MacKenzie, 1966; Kahn, 1974; Cano & Şêrgo, 1991; Paul, 1998; Thackston, 2006a,b). The list in Appendix A also presents various characteristics of the dictionaries, such as target dialects, script, and entry description.

3.1 Before the 20th century

Three major lexicographic resources were published before the 20th century:

Garzoni's Kurdish Grammar and Vocabulary Book (Garzoni, 1787). This dictionary is a part of the earliest scientific European studies on the Kurdish language and civilization which dates back to the late 18th century. The research carried out by various Christian missionaries (Yarshater, 1982). (Garzoni, 1787) collected materials for his *Grammatica e vocabolario della lingua Kurda* (*Grammar and Vocabulary for the Kurdish Language*) (Garzoni, 1787) in Amedi (Amadyia), which is now located in the Kurdistan region of Iraq. This book is an Italian-Kurmanji dictionary and grammar guide which was written to enable missionaries to converse with Kurmanji speakers.

Jaba's Kurmanji Kurdish-French Dictionary (Jaba, 1879). This dictionary presents its entries in both Arabic (Ottoman Turkish script) and Latin orthographies. The latter is used for phonological purposes and therefore can be considered as the pronunciation of the entry. Although definitions and etymological information are mostly provided alongside the entries, the PoS and the gender of the nouns are less frequently present in the dictionary.

Maqdisi's Kurmanji Kurdish-Arabic Dictionary (Mokri, 1987). This dictionary was published in 1892 based on the dialect of Bitlis, now located in Turkey, by a Palestinian Arab Ottoman official. Although neither the PoS nor the gender of nouns is indicated, the present stem of verbs is regularly included. Another version of the dictionary was published with Turkish translations rather than the original Arabic in 1978 (Paşa & Bozarslan, 1978).

3.2 After the 20th century

Kurdish lexicography flourished in the 20th century through the efforts of Kurdish native scholars and orientalists, particularly by the researchers of the former Soviet Union (Leezenberg et al., 2011). This section describes a number of these dictionaries under bilingual, monolingual, and multilingual categories which were published during the mentioned period. These dictionaries are selected based on their contribution significance to Kurdish lexicography.

3.2.1 Bilingual dictionaries

Bakaev's Kurmanji Kurdish-Russian Dictionary (Bakaev, 1957). This dictionary was one of the first linguistic works in the former Soviet Union. The author was a native Russian speaker whose mother tongue was Kurdish. The combination of the author's philological background and his practical knowledge of Kurdish enabled him to produce a standard dictionary. The vocabulary is Kurmanji based on the language of the Kurdish community in the former Armenian Soviet Socialist Republic (SSR) and the former Georgian SSR.

Bakaev collected the dictionary data from various sources, such as folklore texts published mainly during era of the former Soviet Union, the works of the folklorists affiliated with the institutes of the Academy of Sciences of the Armenian SSR and Yerevan State University, the literary work translated into or originally written in Kurdish published in Armenian SSR, and translated textbooks from Russian and Armenian into Kurdish. This explains the presence of many words which were not common in Kurdish daily life (Chyet, 1998).

Kurdoev's Kurmanji Kurdish-Russian Dictionary (Kurdoev, 1960). The author of the dictionary set himself the task to most fully reflect the vocabulary fund of the modern Kurdish language. The vocabulary includes household, agricultural and modern literary language and the press. The dictionary is based on the vocabulary used in a Kurmanji speaking area in Soran, which is currently located in the Kurdistan Region of Iraq. Although the dictionary presents a more diverse vocabulary in comparison to Bakaev's work, the reliability of its data and also its scientific approach have been questioned by some scholars (Chyet, 1998).

Wahbi and Edmonds's Sorani Kurdish-English Dictionary (Wahby & Edmonds, 1966). This dictionary comprises the lexical material of the "standard language of belles-lettres, journalism, official and private correspondence and formal speech as it has been developed, on the basis of the Southern-Kurmanji dialect of Sulaymaniyah in Iraq since 1918" (Mokri, 1987). Moreover, the dictionary contains words unique to the sub-dialects spoken in Erbil, Kirkuk, and Sanandaj. The dictionary does not provide bibliographic information about its lexicographic resources. However, according to Bodrogligeti (1967), Sheikh Mihammadi Khal's Ferhenî Xal by Xal (1960), work by MacKenzie (1961, 1962, 1966), and McCarus (1958) perhaps contributed to the compilation of this dictionary.

Kurdoev and Yusupova's Sorani Kurdish-Russian Dictionary (Kurdoev & Yusupova, 1983). This dictionary is the first Sorani Kurdish-Russian based the Sulaimani sub-dialect of Sorani. The authors compiled the dictionary based on the translations of the entries of dictionaries by Kurdoev (1960) and Mukryani (1950). The information provided for the entries in this dictionary includes pronunciation, PoS, idioms, and expressions.

Chyet's Kurdish-English Dictionary (Chyet & Schwartz, 2003). Chyet's dictionary is a seminal work in Kurdish lexicography containing all the main Kurdish dialects. The entries in Kurmanji Kurdish are in Latin and Arabic orthographies, followed by the PoS, numbered definitions in English, synonyms and variant forms. Moreover, the dictionary contains etymological and linguistic remarks along with expressions and examples with translations in English. Interestingly, relevant forms of a word in Early, Middle and Modern Iranian followed by Sorani, Zaza and Gorani-Hawrami equivalents are provided. Chyet used several dictionaries to compile this resource.

Hakem's Sorani Kurdish-French Dictionary (Hakem, 2012). This dictionary contains

around 22,000 entries, 3,000 variants corresponding to entries, nearly 2,000 sub-entries (compound verbs) and more than 1,000 expressions. There are radicals of each simple verb or that of the compounds when the simple verb is no longer used in the spoken or written language. The grammatical category of each entry is indicated, as well as the language level whenever it seems necessary. The entries are written in Arabic characters and in Latin transcription. In some cases, expressions are also provided for entries. This dictionary focuses on contemporary language in the different registers of writing and speaking, both in the Kurdistan of Iraq and the Kurdistan of Iran.

University of Kurdistan Dictionaries (M. Rohani, 2012, 2018). These two dictionaries were compiled at the University of Kurdistan in Sanandaj based on *Henbane Borîne* (Sharafkandi, 1991) with enriched details added such as pronunciation, etymology, definition, synonyms, translations and variant forms. Moreover, they include neologisms for technical terms. M. Rohani (2012) addressed all Kurdish dialects, which makes it distinctive among bilingual dictionaries.

3.2.2 Monolingual dictionaries

Bedirxan and Keskin (2009) published the first Kurdish-Kurdish dictionary in Kurmanji and later, the *Xal Dictionary* (Xal, 1960) was published as the first Sorani Kurdish monolingual dictionary. In addition, there have been efforts to create dictionaries within Kurdish dialects, such as Habiballah's (Bedar) (2010) dictionary in Hawrami with Sorani translations, Izadpanah's (1978) dictionary in two Southern Kurdish dialects, Laki and Lori, with Sorani translations and Sohrabi and Sreshabadi's (2012) dictionary of the Garusi sub-dialect of Southern Kurdish with Sorani translations. Other monolingual dictionaries which are mostly in Kurmanji Kurdish are Botî (2006), Demîrhan (2007) and Mukryani (2007).

3.2.3 Multilingual dictionaries

Blau's Kurmanji Kurdish-English-French dictionary (Blau, 1965) is the first multilingual Kurdish dictionary which was created based on newspaper articles published in the 1930's and 1940's based on Kurmanji journals. However, the English translations provided in this resource have been questioned by scholars (Chyet, 1998; M., 1966). Several years later, the author published a book consisting of a linguistic analysis, Sorani glossaries and folkloric texts with French translations (Blau, 1975). The two glossaries contain richer descriptions, including gender, part-of-speech, present stem of verbs and oral example texts.

Henbane Borîne (Sharafkandi, 1991) is a Kurdish-Persian dictionary that incorporated all Kurdish dialects in its compilation. In addition to the Persian translations, this resource provides synonyms and senses in Kurdish, including all dialects. Therefore, it sets a foundation for the unification of the dialects. Many dictionaries, which have been compiled following *Henbane Borîne*, have referred to it as one of their essential resources.

3.3 Terminological resources

Various terminological resources exist in Kurdish, such as a glossary of the names of animals (Justi, 1878), glossary of plants (Kasimoğlu, 2013), glossary of law (Talbani, 2006), and engineering (Soĝanci, 2014). A valuable terminological resource for Kurdish is *Kurmancî*, which is a biannual linguistic magazine published by the Kurdish Institute of Paris since 1987. The aim of the magazine is to spread the results of the Institute's linguistic seminars on problems of terminology and standardization of the Kurdish language. The periodical contains headwords in Kurmanji Kurdish and translations in French, English, and Turkish. A database containing the words published in all issues of this periodical is available online[2].

3.4 Electronic dictionaries

Gautier (1996) was the pioneer in creating the first electronic dictionaries for Kurdish in his *Dirêjî Kurdî* (Kurdish Dimension) project. This project aimed at developing a lexicographic software environment specifically for Kurdish to deal with various early-age technical issues such as character representation. FreeDict, as a project which provides open-source bilingual dictionaries for most languages, also has dictionaries in Kurmanji to Turkish, German and English, as well as Sorani to Kurmanji. The dictionaries are publicly available in the TEI XML format. However, the sources of the resources are not clear in all cases. Moreover, there are a few collaboratively created dictionaries, or Wiktionaries[3](available for Kurmanji and Sorani), which provide electronic content. There have been a few efforts in creating electronic lexicons on the Web based on a printed dictionary, but as none of them are documented, we cannot cite them here.

4. Methodology

In order to create our dictionaries, we follow the pipeline illustrated in Figure 3:

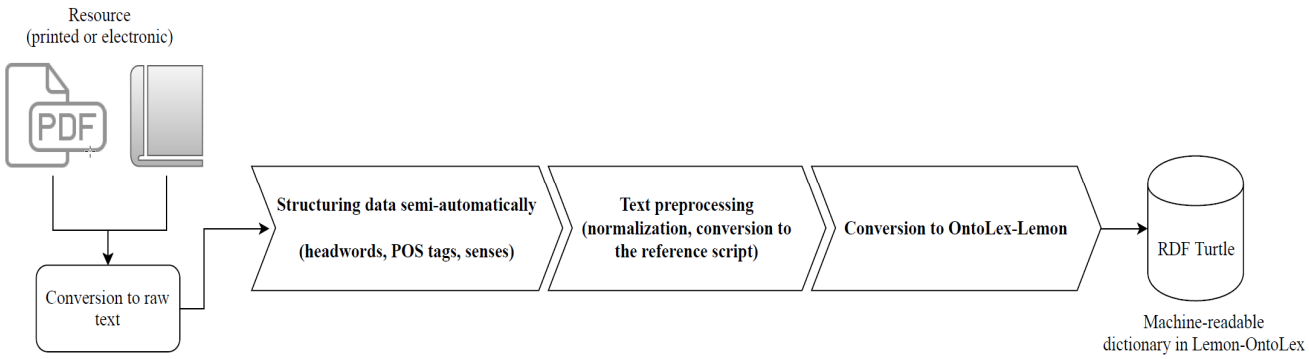

Figure 3: Our resource creation pipeline for creating dictionaries in Ontolex-Lemon from PDF documents.

[2] https://www.institutkurde.org/en/publications/kurmanci/

[3] https://www.wiktionary.org/

4.1 Data collection

As a wide range of published Kurdish dictionaries are available, we selected three dictionaries for our experiment following three selection criteria: i) the number of entries to be manageable in a research project, ii) the availability of the resource and iii) the copyright situation of the resource. Eventually, we selected the word lists provided in three grammar books of Kurmanji (Thackston, 2006a), Sorani (Thackston, 2006b) and Hawrami (MacKenzie, 1966). In addition to reliability, these resources provide a workable sample of a few thousand frequent entries in those dialects. We were not able to find a similar resource for the Kirmashani (Southern Kurdish) dialect.

The Kurmanji and Sorani word lists were available in searchable Portable Document Format (PDF), hence we extracted information into an unstructured text semi-automatically. Because this semi-automatic extraction created some noise, improper transformation to text and misplaced portions of texts, we manually cleaned the text by removing noise and recreating the micro- and macro-structure of the lexicon using tabulations. In the case of the Hawrami lexicon, we had to re-type the word list manually as only the printed book was available.

Moreover, we modified a few traditional lexicographic norms in the resources, such as replacing ~ by headword and placing relevant lexemes of an entry as new entries if with different PoS or etymological roots. Figure 5, on the left, illustrates the Kurdish entry "*bend*" (bond in English) in the Kurmanji-English dictionary where "*~ kirin*" (to arrest, to fetter) and "*man di ~a*" (to wait for) are respectively replaced by "*bend kirin*" and "*man di benda*" as new entries. Similarly, we modified the English translations, particularly in cases where two synonym verbs are provided, the preposition "to" is only provided for the first verb.

Following the data extraction, we unified the orthography and the scripts of the resources. The word lists were originally written in orthographies suggested by the authors and used for teaching purposes. Having various scripts for writing in Kurdish causes a burden for the computation process (Ahmadi, 2019). Moreover, none of the current Kurdish scripts can be used for all Kurdish dialects. Therefore, we suggest a new character setup, illustrated in Figure 1, based on Latin orthography and the phonetics of the language to deal with the missing characters and to accommodate computation needs. The suggested script introduces a single character for the phonemes in all Kurdish dialects, such as ğ and ˆd used in the Zaza and Hawrami dialects, respectively. As the orthographies were based on the phonetics of the language (in Latin), we could automatically transliterate the original text into our suggested orthography. We ignored the Persian-Arabic equivalent of Sorani lexicon at this stage.

4.2 Conversion to OntoLex-Lemon

In recent years there have been efforts to create specific data models providing support

for representing linguistic data on the Semantic Web. The OntoLex-Lemon (McCrae et al., 2017) is a model based on the Lexicon Model for Ontologies (lemon) which provides rich linguistic grounding for ontologies, such as representation of morphological and syntactic properties of lexical entries. This model draws heavily on previous lexical data models, particularly LexInfo (Cimiano et al., 2011), LIR (Montiel-Ponsoda et al., 2008) and LMF (Francopoulo et al., 2006), with improvements such as being RDF-native, descriptive and modular justifying its promise of adaptability in linguistic resource management. The core vocabulary of Lemon is the Ontology Lexicon (Ontolex), known as OntoLex-Lemon, which is illustrated in Figure 4.

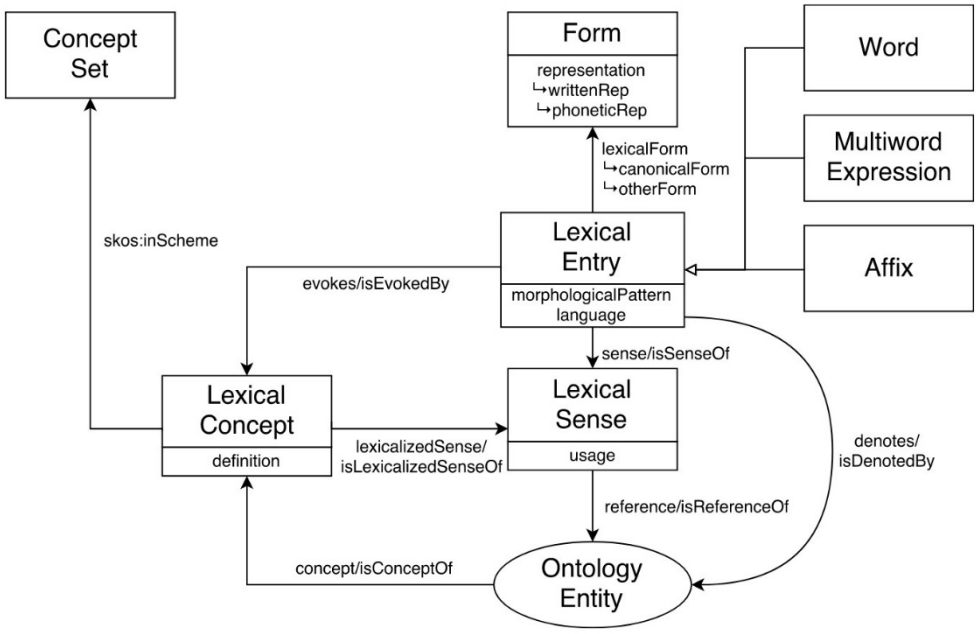

Figure 4: Lemon-OntoLex Core (McCrae et al., 2017).

The previous step yielded a tabular format of the lexicographic information, making it possible to convert the data semi-automatically into RDF triples in OntoLex-Lemon.

Figure 5 illustrates the equivalent of the entry "*bend*" in the Kurmanji-English dictionary in RDF Turtle in Ontolex-Lemon. We have used language tags according to ISO 639-3[4], kmr for Kurmanji, ckb for Sorani and hac for Hawrami (registered as Gorani). As there are many scripts for Kurdish writing, we also include a subtag expressing script following the language tag. For instance, kmr-latn shows that the literal is in Kurmanji Kurdish and written in the Latin script. The script code Arab can be used for Arabic script as well.

[4] https://iso639-3.sil.org/code_tables/639/

bend *f* bond; **li ~a** for the sake of, chained to, waiting for: *divê em êdî li benda sibehê ranewestin* we shouldn't stand around waiting for tomorrow; **~ kirin** *v.t.* to fetter, arrest; **man di ~a** to wait for

```
1  :lexicon a lime:Lexicon;
2      lime:language <www.lexvo.org/page/iso639-3/kmr> ;
3      lime:entry :lex_bend .
4
5  :lex_bend a ontolex:LexicalEntry, ontolex:Word ;
6      ontolex:canonicalForm :form_bend ;
7      rdfs:label "bend"@kmr-latn .
8
9  :form_bend a ontolex:Form ;
10     dct:language <www.lexvo.org/page/iso639-3/kmr> ;
11     ontolex:writtenRep "bend"@kmr-latn ;
12     lexinfo:partOfSpeech lexinfo:noun ;
13     lexinfo:gender lexinfo:feminine ;
14     lexinfo:number lexinfo:singular ;
15     ontolex:sense :bend_n_sense .
16
17 :en_bond a ontolex:LexicalEntry ;
18     dct:language <http://lexvo.org/id/iso639-1/en> ;
19     ontolex:sense :en_bond_sense .
20
21 :trans a vartrans:Translation ;
22    vartrans:source :bend_n_sense ;
23    vartrans:target :en_bond_sense .
24
25 :bend_n_sense a lexicog:UsageExample ;
26    rdf:value "divê em êdî li benda sibehê
          ranewestin."@kmr-latn ;
27    rdf:value "we shouldn't stand around waiting for
          tomorrow."@en .
```

Figure 5: An example entry from our Kurmanji-English dictionary. The original printed entry on the left and the equivalent in RDF Turtle based on the OntoLex-Lemon model.

In addition to OntoLex-Lemon core, we used the following modules:

- Linguistic Metadata (lime) allows to describe metadata at the level of the lexicon-ontology interface with information such as lexical entries and language (lines 1 to 3 in Figure 5).

- Syntax and Semantics (synsem) enables us to describes syntactic behaviour. We use syntactic frames to relate a lexical entry to one of its various syntactic roles, such as the canonical form of the word *bend* described in lines 5 to 7 in Figure 5).

- Lexinfo (lexinfo) (Cimiano et al., 2011) for describing relevant linguistic categories and properties, particularly part-of-speech, gender and number (lines 9 to 15 in Figure 5).

- Variation and translation (vartrans) is used to describe relations between lexical entries, particularly translations. As our resources are not currently connected to any external English resource, we also create entries for English words as shown in lines 17 to 19 in Figure 5.

- Lexicography module (lexicog) (Bosque-Gil et al., 2017) represents information,

structures and annotations commonly found in lexicography. The Lexicographic Resource class in this module is used to represent the original printed entry structures. In addition, we used the UsageExample class for representing examples of the usage of a sense (lines 25 to 27 in Figure 5).

Multi-word expressions (MWEs) are lexical units which are semantically unique, greater than a word, and can bear both idiomatic and compositional meanings (Masini, 2005). Therefore, we create new entries for MWEs using ontolex:MultiwordExpression. Regarding Kurdish MWEs, we could not find any writing standard. In both orthographies, Persian-Arabic-based and Latin-based, words in MWEs are written either with spaces or without. For instance, "*toz-û-telaz*" (dust) can be found as "*tozûtelaz*", "*tozwtelaz*" or "*toz û telaz*" in the literature. Hence, we followed the English norm of using hyphens, i.e. -, for Kurdish MWEs. Furthermore, regarding the idioms, we create new entries as they are semantically different from the canonical forms as well.

5. Analysis

Table 1 provides a statistical analysis of various characteristics of our lexicographic resources. # Entries refers to the number of entries in the electronic dictionary. Put in other terms, it refers to the number of triples with lime:entry properties. This feature does not have the same value as the printed original resources, as idioms and MWEs are presented as new entries in the electronic resources while they are presented in the description of the entries in the printed resources. Furthermore, statistics regarding attributes such as gender, PoS tag, etymological roots, example sentences and idioms are provided. The Sorani and Hawrami dictionaries have the highest number of Gender and PoS tags and etymological roots, respectively.

Resource	Number of entries		Attributes				Polysemy degree
	Word	MWE	Gender & POS	Etymology	# idioms	Examples	
Kurmanji	4172	122	3420 (76.64%)	213 (4.96%)	340	265 (6.35%)	1.03%
Sorani	5683	160	5348 (91.37%)	111 (1.89%)	82	543 (9.55%)	1.06%
Hawrami	1184	165	1184 (87.76%)	242 (17.93%)	123	10 (0.008%)	1.01%

Table 1: Lexicographic resources statistics

We define polysemy degree as the number of unique senses divided by the number of entries. This measure varies in the range of 1.01% and 1.06%, indicating that a small proportion of less than 1% of the entries are polysemous, and for the rest there is only one sense available.

6. Conclusion and future work

In this paper, we provided a review on the current state of Kurdish lexicography and described the development of dictionaries for three out of five main dialects of Kurdish, namely Sorani, Kurmanji and Hawrami. Having more than 60 printed dictionaries and terminological resources, we demonstrate that Kurdish is fairly rich in printed resources, although this is not the case with respect to electronic and machine-readable resources. The lack of such resources makes Kurdish a less-resourced language.

Our lexicographic resources are created using the word lists provided in three grammar books of Kurmanji (Thackston, 2006a), Sorani (Thackston, 2006b) and Hawrami (MacKenzie, 1966) and according to the OntoLex-Lemon model. As Kurdish is written in more than one script and some of the dialectal phonemes do not have a character in those scripts, we suggest a few characters based on the Latin script which can lead to a unification of the scripts. The resources are publicly available for non-commercial use under the CC BY-NC-SA 4.0 license [5] at https://github.com/KurdishBLARK/KurdishLex.

The current study aims at paving the way for Kurdish e-lexicography by developing prototypical resources. Enriching our dictionaries using additional resources and scripts and, linking the dictionaries across dialects and resources, such as KurdNet (Aliabadi et al., 2014), may be addressed in the future work. Creating specific standards for Kurdish, particularly regarding the scripts, will also be suggested as future work. We would also like to highlight solutions to tackle some of the current challenges in Kurdish elexicography such as the following:

- Lexicographic infrastructure: as our findings suggest, more than half of Kurdish dictionaries were created before 2000. In order to create machine-readable version of these resources, retrodigitization tools, such as Optical Character Recognition, are required. On the other hand, tools for creating and maintaining dictionaries are needed.
- Raising awareness: we believe that the lexicography community should be aware of the current computer-based solutions for creating resources and collecting data.
- Creating basic Kurdish text processing tools such as lemmatizer, spell-checker (Salavati & Ahmadi, 2017) and name entity recognizer.
- Copyright issues: the majority of the dictionaries cited in this paper were available online in scanned version or searchable PDF. This is against the copyrights and creator licences, which leads to discouragement in the lexicographers' community.

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

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

8. References

Abdollahpour, H. (2008). *Ferhengî Hejîr: Farsî-Kurdî (Hejîr dictionary: Persian-Kurdish) (Sorani)*, volume 2. Mukryani Publishing House.

Ahmadi, S. (2019). A Rule-based Kurdish Text Transliteration System. *Asian and LowResource Language Information Processing (TALLIP)*, 18(2), pp. 18:1–18:8.

Aliabadi, P., Ahmadi, M. S., Salavati, S. & Esmaili, K. S. (2014). Towards building Kurdnet, the Kurdish Wordnet. In *Proceedings of the Seventh Global Wordnet Conference*. pp. 1–6.

Amin, P. (2003). *Yad dictionary: English-Arabic-Kurdish*. Erbil.

Anter, M. (1967). *Ferhenga Kurdî-Tirkî (Kurdish-Turkish dictionary)*. Istanbul: Yeni Matbaa.

Arif, H. K. (2006). *Govend-Zinar: Ferhengî Farsî-Kurdî (Govend-Zinar Persian-Kurdish Dictionary) (Sorani)*, volume 2. Mukryani Publishing House.

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

Bakaev, C. K. (1957). *Kurdish-Russian dictionary*. State Publishing House of Foreign and National Dictionaries, Moscow.

Bedirxan, C. A. & Keskin, A. (2009). *Ferheng: Kurdî, Kurdî (Kurdish-Kurdish dictionary) (Kurmanji)*, volume 2. Avesta.

Bedirxan, C. A. & Lescot, R. (1970). *Grammaire kurde (Kurdish Grammar)*. Librairie d'Amérique et d'Orient.

Berners-Lee, T., Hendler, J., Lassila, O. et al. (2001). The Semantic Web. *Scientific american*, 284(5), pp. 28–37.

Biemann, C., Heyer, G., Quasthoff, U. & Richter, M. (2007). The Leipzig Corpora Collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, 2007.

Blau, J. (1965). *Dictionnaire kurde-français-anglais (Kurdish-French-English dictionary)*, volume 9. Centre pour l'étude des problèmes du monde musulman contemporain.

Blau, J. (1975). *Le Kurde de Amadiya et de Djabal Sindjar: analyse linguistique, textes folkloriques, glossaires*. Librairie C. Klincksieck.

Bodrogligeti, A. (1967). A Kurdish-English dictionary. By Taufiq Wahby and C. J. Edmonds, pp. xii, 179. Oxford, Clarendon Press, 1966. *Journal of the Royal Asiatic Society of Great Britain & Ireland*, 99(2), pp. 152–155.

Bolelli, N. & Ertekin, N. (2017). Ferhengên Menzûm Di Edebîyata Kurdî De. *Bingöl Üniversitesi Yaşayan Diller Enstitüsü Dergisi*, 3(5), pp. 21–44. (in Kurdish).

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a module for

lexicography in OntoLex. *DICTIONARY News*, 7.

Botî, K. (2006). *Ferhenga Kamêran: Kurdî-Kurdî (Kamêran dictionary: Kurdish-Kurdish) (Kurmanji)*. Spîrêz Press & Publisher. https://books.google.ie/books?id=AVErAQAAIAAJ.

Cano, D. & Şêrgo, M. (1991). *Ferheng Erebî-Kurdî Zaraveyê Kurdmancî (Arabic-Kurdish dictionary) (Kurmanji)*, volume 2. Beirut, Lebanon.

Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer, pp. 7–25.

Chyet, M. (1998). Kurdish Lexicography a Survey and Discussion. *Iran and the Caucasus*, 2(1), pp. 109–118.

Chyet, M. L. & Schwartz, M. (2003). *Kurdish-English Dictionary (Kurmanji)*. Yale University Press.

Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), pp. 29–51.

Darvishian, A. (1997). *Kermanshahi Kurdish Dictionary*. Sahand Publication House.

Demîrhan, U. (2007). Ferhenga Destî–Kurdî bi Kûrdî (Kurdish-Kurdish pocket dictionary).

Ebrahimpour, T. (2008). *Ferhengî Kurdî-Îngilîsî (Kurdish-English Dictionary) (Sorani)*. Saha Publications, Tehran.

Ertekin, Z. (2017). İlk Manzum Türkçe-Kürtçe Sözlük: Nûbihara Mezinan. *e-Şarkiyat İlmi Araştırmaları Dergisi/Journal of Oriental Scientific Research (JOSR)*, 9(1), pp. 89–105.

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

Farizov, I. (1957). *Russian-Kurdish dictionary (Kurmanji)*. State Publishing House of Foreign and National Dictionaries, Moscow.

Forcada, M. L., Esplà-Gomis, M., Pérez-Ortiz, J. A., Sánchez-Cartagena, V. M. & SánchezMartínez, F. (2019). D1.1 – Survey of relevant low-resource languages. Technical report, Global Under-Resourced MEdia Translation (GoURMET).

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*, p. 5.

Garzoni, M. (1787). *Grammatica e vocabolario della lingua kurda composti dal p. Maurizio Garzoni de'predicatori ex-missionario apostolico*. nella Stamperia della Sacra Congregazione di Propaganda Fide.

Gautier, G. (1996). Dirêjî Kurdî: a lexicographic environment for Kurdish language using 4th Dimension R . In *5th International Conference and Exhibition on Multilingual Computing (ICEMCO)*, volume 5. pp. Session–of.

Gautier, G. (1998). Building a Kurdish Language Corpus: An Overview of the

Technical Problems. *Proceedings of ICEMCO*.

Gewranî, A. S. A. (1985). *Ferhenga Kurdî Nûjen: Kurdî-Erebî (Nûjen Kurdish dictionary: Kurdish-Arabic) (Kurmanji)*. Amman: A.S.A. https://books.google.ie/books?id=kJncNQAACAAJ.

Gharib, K. (1975). *Arabic-English-Kurdish Dictionary*. Alajial, Baghdad.

Gökırmak, M. & Tyers, F. M. (2017). A dependency treebank for Kurmanji Kurdish. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 64–72.

Habiballah (Bedar), J. (2010). *Wişename (Lexicon)*. Aras Publishing and Printing House.

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

Hakem, H. (2012). *Dictionnaire kurde-français: Sorani (Kurdish-French dictionary: Sorani)*. L'Asiathèque, Maison des langues du monde.

Hassani, H. & Medjedovic, D. (2016). Automatic Kurdish dialects identification. *Computer Science & Information Technology*, 6(2), pp. 61–78.

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

Hassanpour, A. (1992). *Nationalism and language in Kurdistan, 1918-1985*. Edwin Mellen Press.

Ismail Hassan, S. (2019). *Kurdish-English Dictionary*. Tafseer Office.

Izadpanah, H. (1978). *Lak and Lor dictionary*. Kurdish Scientific Council, Baghdad.

Izoli, D. (1992). *Ferheng: Kurdi-Tirki, Türkçe-Kürtçe (Dictionary: Kurdish-Turkish, Turkish-Kurdish) (Kurmanji)*. Istanbul: Deng Yayinlari.

Jaba, A. (1879). *Dictionnaire kurde-français (Kurdish-French dictionary) (Kurmanji)*. Commissionaire de l'Académie Impériale des Sciences.

Jalilian, A. (2010). *Ferhengî Başûr (Başûr Dictionary)*. Aras Publishing and Printing House.

Jalilian, A. A. (2009). *Ferhengî Başûr (Başûr Dictionary) (Southern Kurdish)*. Enstîtûy Kelepûrî Kurdî.

Justi, F. (1878). *Les noms d'animaux en kurde (name of animals in Kurdish) (Kurmanji)*. Imprimerie nationale.

Kahn, M. (1974). *Kurmanji-English, Enlgish-Kurmanji Lexicon*. Ann Arbor: The University of Michigan. Unpublished manuscript.

Karadaghi, R. (2006). *The Azadi: English-Kurdish Dictionary*. Ehsan Publication House.

Kasımoğlu, A. (2013). *Ferhenga Naven Nebatan A Kurdi (Dictionary of Plants in Kurdish, Turkish and Latin) (Kurmanji)*. İmaj Matbaacılık Sanayi. https://books.google.ie/books?id=4p0JugEACAAJ.

Keidane, K., Mukriani, K. & Mitrokhina, V. (1977). *Educational Russian-Kurdish Dictionary*. Moscow, Publisher "Russian Language".

Khalidgul, M. (2002). *Ferhenga Gulî: Farsî-Kurdî (Gulî dictionary: Persian-Kurdish) (Kurmanji)*. Spîrêz Press & Publisher.

Kiani Kolivand, K. (2011). *Kian dictionary: Laki lexicon (in Persian)*, volume 2. Sifa, Khorramabad.

Klyne, G. & Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. https://www.w3.org/TR/rdf-concepts/.

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

Kurdoev, K. K. & Yusupova, Z. (1983). *Kurdish-Russian Dictionary (Sorani).* Moscow, Publisher "Russian Language".

Kurdoev, K. K. (1960). *Kurdish-Russian dictionary (Kurmanji).* State Publishing House of Foreign and National Dictionaries, Moscow.

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

MacKenzie, D. N. (1961). The origins of Kurdish. *Transactions of the Philological Society*, 60(1), pp. 68–86.

MacKenzie, D. N. (1962). *Kurdish dialect: studies*, volume 2. Oxford University Press.

MacKenzie, D. N. (1966). Joyce Blau: Kurdish-French-English dictionary. *Bulletin of the School of Oriental and African Studies*, 29(3), p. 672–673.

MacKenzie, D. N. (1966). *The dialect of Awroman (Hawraman-i Luhon) Grammatical sketch, texts, and vocabulary.* Copenhagen: Munksgaard.

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

Masini, F. (2005). Multi-word expressions between syntax and the lexicon: the case of Italian verb-particle constructions. *SKY Journal of Linguistics*, 18(2005), pp. 145–173.

Mayi, T. M. (2009). *Ferhenga Mayî: Kurdî-Erebî (Mayî Dictionary: Kurdish-Arabic) (Kurmanji).* Spîrêz Press & Publisher.

McCarus, E. N. (1958). *A Kurdish Grammar: descriptive analysis of the Kurdish of Sulaimaniya, Iraq.* 10. American Council of Learned Societies. http://files.eric.ed.gov/fulltext/ED089545.pdf.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntolexLemon model: development and applications. In I. Kosem et al. (eds.) *Proceedings of eLex 2017 conference, Leiden, Netherlands.* Brno: Lexical Computing, pp. 19–21.

Miller, G.A . (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Mokri, M. (1987). *Dictionnaire Kurde-Arabe de Dia'Ad-din Pacha Al-Khalidi.* Kanz almutun wa-al-dirasat al-madhhabiyah, al-lughawiyah wa-al-ijtima'iyah, al-hadarah alIslamiyah, al-lisan wa-al-thaqafah al-Iraniyah. Libraire du Liban. https://books.google.ie/books?id=JrYpDgAAQBAJ.

Montiel-Ponsoda, E., De Cea, G. A., Gómez-Pérez, A. & Peters, W. (2008). Modelling multilinguality in ontologies. *Coling 2008: Companion volume: Posters*, pp. 67–70.

Mukryani, G. (1950). *Rêber dictionary: Arabic-Kurdish (Sorani).* Erbil.

Mukryani, G. (1961). *Mahabad dictionary: Kurdish-Arabic (Sorani)*. Erbil.

Mukryani, G. (1966). *Pellke-zêrrîne dictionary: Kurdish-Arabic-Persian-French-English (Sorani)*. Erbil.

Mukryani, G. (2005). *Nobere dictionary: Arabic-Kurdish dictionary for Educational Puropses (Sorani)*. Aras Publishing House, Erbil.

Mukryani, K. (2007). *Haraşan dictionary: Kurdish-Kurdish (Sorani)*. Kurdistan region, Ministry of Culture.

Nahid, M. (2011). *Ferhengî Nahîd: Kurdî-Kurdî-Farsî (Nahid dictionary: Kurdish-KurdishPersian) (Sorani)*. Akadîmyay Kurdî, Erbil.

Nanwazade, A. (2005). *Ferhengî Kurdîy Herman (Herman Kurdish Dictionary) (Sorani Kurdish)*, volume 2. Mukryani Publishing House.

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.

Nawkhosh, S. (2012). *Kurdish-Arabic-English Dictionary*.

Nizameddin, F. (2003). *Ferhengî Estêregeşe: Kurdî-Erebî (Estêregeşe dictionary: KurdishArabic) (Sorani)*. Aras, Erbil.

Nodeyî, M. (1936). *Ahmadi dictionary: Arabic-Kurdish*. Jyan Publishing house, Sulaymaniyah.

Özcan, M. (1997). *Zazaca-Türkçe Sözlük (Zaza-Turkish dictionary)*. Kaynak Yayınları.

Paşa, Y. Z. & Bozarslan, M. E. (1978). *Kürtçe-Türkçe sözlük (Kurdish-Turkish dictionary)*. Çıra Yayınları.

Paul, L. (1998). *Zazaki: Grammatik und Versuch einer Dialektologie (Grammar and dialectology experiment)*. Reichert Verlag.

Qazzaz, S. (2000). *The Sharezoor Kurdish-English dictionary: Farhang-i Sharazur kurdiinglizi*. Aras Publishing and Printing House.

Rizgar, B. (1993). *Kurdish-English, English-Kurdish Dictionary (Kurmanji)*. MF Onen, London.

Rohani, M. (2012). *University of Kurdistan Dictionary: Persian-Kurdish*, volume 3. University of Kurdistan, Sanandaj Iran.

Rohani, M. (2018). *University of Kurdistan Dictionary: Kurdish-Kurdish-Persian*, volume 4. University of Kurdistan, Sanandaj Iran.

Saadalla, S. (1998). *Saladin's English-Kurdish Dictionary (Kurmanji)*. Sweden. (Arabic script).

Saadalla, S. (2000). *Saladin's English-Kurdish Dictionary (Kurmanji)*. Avesta. (Latin script).

Salavati, S. & Ahmadi, S. (2017). Building a Lemmatizer and a Spell-checker for Sorani Kurdish. *LTC'17 The 8th Language & Technology Conference*.

Selma Abdallah, K. A. (2006). *English-Kurdish Kurdish-English Dictionary*. Star Publications (P) Ltd., India.

Sharafkandi, A. H. (1991). *Hanbana Borina: Kurdish-Persian dictionary (Sorani)*, volume 2. Soroush, Tehran.

Shi, L. & Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet

and WordNet for robust semantic parsing. In *International conference on intelligent text processing and computational linguistics*. Springer, pp. 100–111.

Siabandov, S. & Châchân, A. (1957). *Armenian-Kurdish Dictionary*. State Press of Armenia (HayPetHrat), Yerevan.

Soğanci, M. (2014). *Ferhenga Zaravên Teknîki: Kurdî-Türkçe-English (Dictionary of Technical Terms: Kurdish-Turkish-English) (Kurmanji)*. Türkiye Mühendis ve Mimar Odaları Birliği.

Sohrabi, R. & Sreshabadi, J. (2012). *Farhang-e Garus (Garus Dictionary) (Southern Kurdish)*. University of Kurdistan.

Talbani, N. (2006). *Legal dictionary: Arabic-Kurdish-French-English (Sorani)*. Hoshyari Publication House.

Thackston, W. M. (2006a). *Kurmanji Kurdish:-A Reference Grammar with Selected Readings*. Harvard University.

Thackston, W. M. (2006b). *Sorani Kurdish–A Reference Grammar with Selected Readings*. Harvard University.

Turgut, H. (2001). *Zazaca-türkçe sözlük (Zaza-Turkish dictionary)*. Wêjiayişê Tiji-Tij Yayınları.

Turgut, H. (2008). *Türkçe-Zazaca Sözlük (Turkish-Zaza dictionary)*. Do Yayınları.

Ulumaskan, A. (2016). *Ferheng - Wörterbuch: Kurdî - Almanî, Kurdisch - Deutsch & Deutsch - Kurdisch, Almanî - Kurdî (Kurdish-German, German-Kurdish dictionary) (Kurmanji)*. Mezopotamien Verlag und Vertrieb GmbH. https://books.google.ie/books?id=LK9YnQAACAAJ.

Wahby, T. & Edmonds, C. J. (1966). *A Kurdish-English Dictionary*. Clarendon Press.

Walther, G. & Sagot, B. (2010). 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)*, p. 8.

Walther, G., Sagot, B. & Fort, K. (2010). Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *International conference on lexis and grammar*, p. 0.

Xal, M. (1960). *Ferhengî Xall (Xall dictionary) (Sorani)*. Aras Publishing House, Erbil. Yarshater, E. (1982). *Encyclopaedia Iranica*, volume 2. Routledge & Kegan Paul.

Yaseen, R. & Hassani, H. (2018). Kurdish Optical Character Recognition. *UKH Journal of Science and Engineering*, 2(1), pp. 18–27.

Yıldırım, K. (2008). *Nubihara Biçûkan, Arapça-Kürtçe-Arapça Sözlük*. Avesta publishing house.

Zîlan, R. (1989). *Ferhenga Swêdî-Kurdi (Kurmancî) (Swedish-Kurdish dictionary) (Kurmanji)*. SIL, Statens Institut för Läromedel. https://books.google.ie/books?id=WHX2AQAACAAJ.

A. Appendix

There are reportedly more than 71 dictionaries and terminological resources available for Kurdish (Jalilian, 2010). In the following list, however, we only provide those to which we could have access to. In order to save space, we used the following symbols: † to refer to an unsystematic script which is not based on the known orthographies, ‡ to denote Ottoman Turkish Arabic script, * to show our estimation based on the number of the pages and density of entries per page. Furthermore, Sor., Kur., SK, HK are used to respectively refer to Sorani, Kurmanji, Southern Kurdish (Kirmashani) and Hawrami dialects. In the Script column, P-A (Persian-Arabic), L (Latin) and C (Cyrillic) are used to show the scripts in which the Kurdish entries or lexemes are written. → and ↔ are used to show translation directions from source language (in the left) to the target language (in the right). In cases of uncertainty, we use ?.

Table 2: The list of Kurdish dictionaries in chronological order based on which the Kurdish lexicography review in Section 3 is carried out in this paper

No	Author	Type	Year	Languages	Entries	Script	Description
1	(Garzoni, 1787)	bilingual	1787	Italian→Kur.	5,250*	L†	translation
2	(Jaba, 1879)	bilingual	1879	Kur.→French	14,340*	P-A‡ and L†	translation, example sentences
3	(Mokri, 1987)	bilingual	1892	Kur.→Arabic	7,200*	P-A‡	translations, present stem of verbs
4	(Mukryani, 1950)	bilingual	1950	Arabic→Sor.	15,000	P-A	translations
5	(Bakaev, 1957)	bilingual	1957	Kur.→Russian	14,000	C	translations, gender, expressions, variant forms
6	(Farizov, 1957)	bilingual	1957	Russian→Kur.	30,000	L	translations, gender, expressions, variant forms
7	(Siabandov & Châchân, 1957)	bilingual	1957	Armenian→Kur.	23,000	C	translation
8	(Kurdoev, 1960)	bilingual	1960	Kur.→Russian	34,000	C	detailed translations with polysemy, gender, expressions, variant forms

9	(Xal, 1960)	monolingual	1960	Sor.→Sor.	22,000*	P-A	definition, synonyms
10	(Mukryani, 1961)	bilingual	1961	Sor.→Arabic	30,000	P-A	translations
11	(Bedirxan & Keskin, 2009)	monolingual	1962	Kur.→Kur.	15,000*	L	synonyms
12	(Blau, 1965)	multilingual	1965	Kur.→(French-English)	6,000*	L	translations, gender, present stem of verbs, limited PoS
13	(Wahby & Edmonds, 1966)	bilingual	1966	Sor.→English	6,500*	L†, P-A	PoS, synonyms and variant forms, rich description
14	(Mukryani, 1966)	multilingual	1966	Sor.→(PersianArabic-English-French)	4,000*	P-A	translations
15	(MacKenzie, 1966)	bilingual	1966	HK→English	1,000	L†	translation, PoS, gender, idioms, example sentences, variant forms
16	(Anter, 1967)	bilingual	1967	Kur.→Turkish	?	L	simple translations
17	(Blau, 1975)	bilingual	1975	(Kur.Sor.)→French and English	2,000*	L	translation, gender, PoS
18	(Gharib, 1975)	multilingual	1975	Arabic→Sor. and English,	1,000*	P-A and L	illustrations
19	(Keidane et al., 1977)	bilingual	1977	Sor.→Russian	2,100	P-A	translation
20	(Paşa & Bozarslan, 1978)	bilingual	1978	Kur.→Turkish	7,200*	L	simple translations, present stem of verbs
21	(Izad-panah, 1978)	multilingual	1978	(Lori- Laki) ↔ (Sor.-Persian)	4,800*	P-A	translations

22	(Kurdoev & Yusupova, 1983)	bilingual	1983	Sor.→Russian	25,000	P-A	translations, PoS, gender, expressions, variant forms
23	(Gewranî, 1985)	bilingual	1985	Kur.→Turkish	25,000*	L	translations
24	(Zîlan, 1989)	bilingual	1989	Swedish→Kur.	5,000	L	translations, synonyms, illustrations
25	(Sharafkandi , 1991)	multilingual	1991	Sor.→(Sor.-Persian)	60,000	P-A†	translations, synonyms
26	(Izoli, 1992)	bilingual	1992	Turkish↔Kur.	25,000-30,000	L	translations, definitions, gender, PoS
27	(Rizgar, 1993)	bilingual	1993	Kur.↔English	15,000	L	translations, PoS, gender, synonyms, expressions, variant forms
28	(Darvishian, 1997)	bilingual	1997	SK→Persian	?	P-A	translation, pronunciation
29	(Özcan, 1997)	bilingual	1997	Zaza→Turkish		L	
30	(Saadalla, 1998)	bilingual	1998	English→Kur.	72,000	P-A	translation, gender
31	(Qazzaz, 2000)	bilingual	2000	Sor.→English	10,000*	P-A and L	translation, synonyms, PoS, idioms, proverbs
32	(Saadalla, 2000)	bilingual	2000	English→Kur.	72,000	L	translation, PoS
33	(Turgut, 2001)	bilingual	2001	Zaza→Turkish	?	L	?
34	(Khalidgul, 2002)	bilingual	2002	Persian→Kur.	4,000	P-A	translations
35	(Chyet & Schwartz, 2003)	bilingual	2003	Kur.→English	59,360*	P-A and L	translations, PoS, gender, expressions, synonyms, variant forms, etymology, example sentence

36	(Demîrhan, 2007)	monolingual	2003	Kur.→Kur.	19,680*	L	?
37	(Nizameddin, 2003)	bilingual	2003	Sor.→Arabic	13,650*	P-A	translations, synonyms
38	(M. Amin, 2003)	multilingual	2003	English→(Sor.-P-A)	35,000*	P-A	translation
39	(Mukryani, 2005)	bilingual	2005	Arabic→Sor.	25,000	P-A	translations
40	(Nanwazade, 2005)	monolingual	2005	Sor.→Sor.	10,000*	P-A	definitions, etymology, idioms
41	(Botî, 2006)	monolingual	2006	Kur.→Kur.	15,000*	L	gender, definition, synonyms
42	(Arif, 2006)	bilingual	2006	Persian→Sor.	36,300	P-A and L	translations
43	(Selma Abdallah, 2006)	bilingual	2006	Sor.↔English.	3,300*	P-A and L	translations
44	(Karadaghi, 2006)	bilingual	2006	English→Sor.	44,000	L and P-A	translations
45	(Mukryani, 2007)	monolingual	2007	Sor.→Sor.	3,000*	P-A	?
46	(Abdollahpour, 2008)	bilingual	2008	Persian→Sor.	28,000	P-A	translations, synonyms
47	(Ebrahimpour, 2008)	bilingual	2008	Sor.→English	40,800*	P-A and L	translation
48	(Turgut, 2008)	bilingual	2008	Turkish→Zaza	?	L	?
49	(Jalilian, 2009)	multilingual	2009	SK→(Sor.-Persian)	31,600	P-A and L	translations, example sentences
50	(Habiballah (Bedar), 2010)	monolingual	2009	HK→Sor.	56,000*	P-A	synonyms
51	(Mayi, 2009)	bilingual	2009	Kur.→Arabic	20,700*	L and P-A	translation, definitions

52	(Ismail Hassan, 2019)	bilingual	2009	Sor.→English	35,000*	P-A and L	synonyms, PoS, pronunciation
53	(Kiani Kolivand, 2011)	bilingual	2011	Laki→Persian	30,000	?	translations, etymology
54	(Nahid, 2011)	multilingual	2011	Sor.→(Sor.-Persian)	21,000*	P-A	translation
55	(Hakem, 2012)	bilingual	2012	Sor.→French	?	P-A	?
56	(Nawkhosh, 2012)	bilingual	2012	Sor.→Arabic and English	3,000*	P-A and L	synonyms
57	(M. Rohani, 2012)	bilingual	2012	Persian→Sor.	18,500	P-A	translation, pronunciation, PoS, idioms, example sentences, variant forms, synonyms
58	(Sohrabi & Sreshabadi, 2012)	multilingual	2012	Garusi→(Sor.-Persian)	8,000	P-A	translation, pronunciation
59	(Ulumaskan , 2016)	bilingual	2016	Kur.↔German	25,000	L	?
60	(M. Rohani, 2018)	multilingual	2018	Sor.→(Sor.-Persian)	93,000	P-A	translation, PoS, idioms, example sentences, variant forms