Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus, Digital Scholarship in the Humanities.

Article *in* Literary and Linguistic Computing · February 2019

DOI: 10.1093/llc/fqy074

Aran Emînî
University of Avignon
27 PUBLICATIONS 201 CITATIONS
SEE PROFILE

Hadi Veisi
University of Tehran
107 PUBLICATIONS 1,295 CITATIONS
SEE PROFILE

Hawre Hosseini
Toronto Metropolitan University
20 PUBLICATIONS 137 CITATIONS
SEE PROFILE

Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus

Hadi Veisi
Faculty of New Sciences and Technologies, University of Tehran, Iran

Mohammad MohammadAmini
IT Department, Faculty of Engineering, Tarbiat Modares University, Iran

Hawre Hosseini
Electrical and Computer Engineering, Ryerson University, Canada

Abstract

In this article, we introduce the first Kurdish text corpus for Central Kurdish (Sorani) branch, called AsoSoft text corpus. Kurdish language, which is spoken by more than 30 million people, has various dialects. As one of the two main branches of Kurdish, Central Kurdish is the formal dialect of Kurdish literature. AsoSoft text corpus is of size 188 million tokens and has been collected mostly from Web sites, published books, and magazines. The corpus has been normalized and converted into Text Encoding Initiative XML format. In both collecting and processing the text, we have faced several challenges and have proposed solutions to them. About 22% of the corpus is topic annotated with six topic tags, and a topic identification task has been done to evaluate the correctness of annotation. The computational experiments of the Central Kurdish text processing are also presented with the support of related supplementary statistics. For the first time, the validity of Zipf's law for Central Kurdish is presented and also perplexity of this language is calculated using standard N-gram language models. The perplexity of Central Kurdish is 276 for a tri-gram language model.

Correspondence:
Hadi Veisi, University of Tehran, Faculty of New Sciences and Technologies (FNST), North Kargar St., Tehran, Iran.
E-mail:
h.veisi@ut.ac.ir

1 Introduction

Kurdish language is a member of the Indo-Iranian branch of Indo-European languages which is spoken by more than 30 million people in Western Asia, mainly in Iraq, Turkey, Iran, Syria, Armenia, and Azerbaijan. The Kurdish language has a variety of dialects and owns its own grammatical system and rich vocabularies (Kaveh, 2005; Rokhzadi, 2011). The two most widely spoken dialects of Kurdish are Central Kurdish (also called Sorani) and Northern Kurdish (also called Kurmanji). Other dialects spoken by smaller populations are Zazaki and Gorani (also known as Hawrami).

doi:10.1093/llc/fqy074

Although spoken by a large population, Kurdish suffers from the unavailability of enough resources for its computational processing purposes. Text corpora, which contain sample texts of a language, are widely used in a variety of natural language processing (NLP) applications. A text corpus, as an essential language resource, can be used in both spoken and written language processing applications such as speech recognition (for language modeling and lexicon extraction) and information retrieval systems (Wynne, 2005).

Up to now, there have been only a few attempts to prepare the linguistic resources for Kurdish, and therefore, only limited research studies have been done on Kurdish language processing (Gautier, 1998; Geoffrey, 2002; Barkhoda *et al.*, 2009; Walther and Sagot, 2010; Walther *et al.*, 2010; Walther, 2011; SheykhEsmaili and Salavati, 2013). In the work by Gautier (1998), the necessity and challenges of building a Kurdish language text corpus have been addressed. Geoffrey (2002) collected a corpus called corpus of contemporary Kurdish newspaper texts, which is a very small text collection including only 214,000 words of Kurmanji dialect collected from *Azadya Welat* newspaper and CTV broadcasting company. Walther and Sagot (2010) introduced a three-step semi-supervised methodology for developing a morphological lexicon for less-resourced languages and applied it for Central Kurdish. They have extracted a lexicon from a small corpus collected from the blog of the program Ruwange broadcasted by the Kurdish channel Roj TV (Ruwange Blog, 2015). The corpus size contained 590,568 tokens and 62,993 unique types. A similar research has been done by the same team for Northern Kurdish (Walther *et al.*, 2010). The aim of the research studies by Walther and Sagot (2010) and (Walther *et al.* (2010) was to build an annotated lexicon and not a corpus; therefore, those research studies did not provide a text corpus for the Kurdish language. (SheykhEsmaili and Salavati (2013) introduced Pewan text corpus for two dialects of Kurdish language, Sorani and Kurmanji, and have used it as a test set for information retrieval applications (SheykhEsmaili *et al.*, 2013). Pewan corpus was collected from the articles dated between 2003 and 2012 from two online news agencies, Peyamner (Peyamner, 2015) and Sorani Kurdish Web site of Voice of America (VOA, 2015). The size of Pewan corpus for Sorani dialect is about 18 million words and for Kurmanji dialect is about 4 million words (SheykhEsmaili and Salavati, 2013). In another research for preparing computational linguistic resources of Sorani Kurdish (Hosseini *et al.*, 2015), a generative lexicon including 35,000 tokens has been collected.

Despite the existence of the mentioned resources, the Kurdish language requires a relatively large-scale text corpus to be used in real-world language processing applications by providing necessary information. Accordingly, in this article, we introduce the first Kurdish text corpus, AsoSoft text corpus,[1] for Central Kurdish. The corpus contains about 188 million words and is collected from different sources including Kurdish Web sites, published books, magazines, and newspapers. While preparing this corpus, we faced various difficulties and gained valuable experiences. Examples of the challenges are the limited number of sources, collecting data from various sources having different and non-standard encodings, mixing Kurdish documents with Arabic and Persian documents, various miss-spelling problems and several normalization difficulties. The details of these challenges and the proposed methods for handling them are discussed in the article. This corpus will be a source to do NLP research in Kurdish, such as statistical language modeling using N-gram or neural network methods (which are used in various applications such as speech recognition, machine translation, and spell checkers), training word embedding representation, lexicon extraction based on the most frequent words, topic identification, language identification, and other linguistics analysis.

In the continuant in this article, we first overview the Kurdish language focusing on Central Kurdish in Section 2. In Section 3, the details of the AsoSoft text corpus are described, including the collection, the specifications of the corpus, and the most important challenges in the processing and normalization of the corpus. Finally, the conclusion and future works are given in Section 4. For the sake of readability, we use Kurdish instead of Central Kurdish (Sorani) where there is no need to make a distinction between Kurdish dialects.

2 Kurdish Language

Kurdish, a language spoken by more than 30 million people in Western Asia, has many dialects, with Northern (Kurmanji) and Central (Sorani) being the two major ones. Kurmanji is spoken in the northern areas of Kurdistan (in Turkey, Syria, and northern Iraq) and is written in Latin (Roman) script; Sorani dialect is spoken mainly in southeastern regions, including Iran and Iraq, and is mostly written in a customized version of the Arabic script. Although Kurmanji has a greater number of speakers than Sorani, the latter one is more standardized and has more written resources due to it being the formal dialect of Kurdish literature.

The Arabic-based writing system for Central Kurdish was first established in the 1920s (Nebez, 1993) and has undergone drastic changes ever since on different occasions. Central Kurdish has thirty-four letters, as shown in Table 1. In this table, letters' different writing forms (joiner and non-joiner) with an example for each and also their corresponding phonemes are given. Kurdish is a phonemic language, i.e. each letter corresponds almost to one phoneme. However, there are some exceptions. As given in Row 28, the letter 'ی' is both pronounced as /j/ (a glide consonant) and as /i/ (a vowel). Also, as in Row 32, the letter 'و' is pronounced both as /w/ (a glide consonant) and as /ʊ/ (a vowel). This letter is also written in a repeated form as 'وو' and

Table 1 Central Kurdish letters

No	Symbol	Letter form				Example				Phoneme (IPA)	Unicode
		Isolated (non-joiner)	Joiner			Isolated (non-joiner)	Joiner				
			Initial	Medial	Final		Initial	Medial	Final		
1	ā (hamza)	ء	ئـ	ـئـ	–	–	ئێواره	مەسئوول	–	ʔ	0626
2	b	ب	بـ	ـبـ	ـب	ئەدەب	باوان	دڵبەر	غەریب	b	0628
3	p	پ	پـ	ـپـ	ـپ	گڵۆپ	پێنووس	پسپۆڕ	کۆسپ	p	067E
4	t	ت	تـ	ـتـ	ـت	دەسەڵات	تەسک	دەستپێک	ڕاست	t	062A
5	j	ج	جـ	ـجـ	ـج	گۆج	جێگە	بێجگە	پێنج	d͡ʒ	062C
6	ch	چ	چـ	ـچـ	ـچ	ورچ	چەم	بچکۆلە	پێچ	t͡ʃ	0686
7	ḥ	ح	حـ	ـحـ	ـح	ڕۆح	حۆل	زەحمەت	مەسیح	ħ	062D
8	kh	خ	خـ	ـخـ	ـخ	بەرخ	خاوێن	تەرخان	شێخ	x	062E
9	d	د	–	ـد	ـد	کورد	–	بێدین	گوند	d	062F
10	r	ر	–	ـر	ـر	کار	–	برین	تێر	ɾ	0631
11	rr	ڕ	–	ـڕ	ـڕ	گۆڕین	–	بڕیار	پڕ	r	0695
12	z	ز	–	ـز	ـز	بەرز	–	پارێزگا	ڕێز	z	0632
13	zh	ژ	–	ـژ	ـژ	کەژ	–	بژار	قژ	ʒ	0698
14	s	س	سـ	ـسـ	ـس	کاس	سوور	بسک	پیس	s	0633
15	sh	ش	شـ	ـشـ	ـش	ڕەش	شانە	نیشتمان	ئێش	ʃ	0634
16	ẹ	ع	عـ	ـعـ	ـع	زەبەلۆع	عەلەشیش	بێعار	واقیع	ʕ	0639
17	gh	غ	غـ	ـغـ	ـغ	ساغ	غەوارە	داڵغە	قەرەباڵغ	G	063A
18	f	ف	فـ	ـفـ	ـف	ماف	فرە	ئەلفبێ	یونیسێف	f	0641
19	v	ڤ	ڤـ	ـڤـ	ـڤ	سۆلاڤ	ڤیان	بڤە	پەیڤ	v	06A4
20	q	ق	قـ	ـقـ	ـق	لاق	قامیش	خوڵقێنەر	بڵق	q	0642
21	k	ک	کـ	ـکـ	ـک	بووک	کڕ	دیکە	ڕێک	k	06A9
22	g	گ	گـ	ـگـ	ـگ	ورگ	گەمارۆ	هەنگاو	درەنگ	g	06AF
23	l	ل	لـ	ـلـ	ـل	گەل	لار	شیلان	چل	l	0644
24	ll	ڵ	–	ـڵـ	ـڵ	هەوڵ	–	سکاڵا	لێڵ	ɫ	06B5
25	m	م	مـ	ـمـ	ـم	کەم	مان	هێما	دێم	m	0645
26	n	ن	نـ	ـنـ	ـن	کۆن	نەوا	بنەما	بن	n	0646
27	h	هـ	هـ	ـهـ	ه	مەهـ	هەناسە	نهێنی	بێهـ	h	06BE
28	y	ی	یـ	ـیـ	ـی	کەی	یار	پرسیار	کوێی	j	06CC
28	I	ی	–	ـیـ	ـی	نەوی	–	قامیش	سی	i	06CC
29	A	ا	–	ا	ـا	ئاوا	–	باران	گەرما	ɑ	0627
30	e	ێ	–	ـێـ	ـێ	ئەوێ	–	ئێوارە	پێ	e	06CE
31	o	ۆ	–	ـۆ	ـۆ	زۆر	–	بۆر	شەوبۆ	o	06C6
32	w	و	–	ـو	ـو	چەوت	–	گوێز	چێو	w	0648
32	u	و	–	ـو	–	کاروبار	–	کورد	–	ʊ	0648
32	uu	وو	–	ـو	ـو	زوو	–	لووت	شوو	u	–
33	a	ە	–	ـە	ـە	گەورە	–	بەش	بەڵگە	a	06D5
34	i	–	–	–	–	–	–	دڵ	–	ɪ	–

is pronounced as /u/ (a long vowel). It is to be emphasized that sometimes letter 'وو' is indicated as an independent letter, for example in the word 'سوور' (meaning red). However, it is in fact a repetition of letter 'و' /ʊ/ and is also not considered as an independent letter in the last released keyboard for Kurdish by the Department of Information Technology of Kurdistan Regional Government (DepIT-KRG, 2015).

As the one-to-one mapping between the letters and the phonemes in Table 1 indicates, Kurdish is a phonographic language. However, 'ی' and 'و' are the exceptions resulting in a one-to-two mapping. Fortunately, there are linguistic rules to distinguish the different pronunciations of those letters (Ilkhanizadeh, 2006): if they appear as the second letter of a syllable, they will be in their vowel forms (i.e. /ʊ/ for 'و' and /i/ for 'ی'); if these letters occur in other positions in a syllable, they will be in their consonant glide forms. Although it is not always easy to determine syllable boundaries in the text, the mentioned rules could be used in some cases in which syllable boundary detection proves to be easier, such as the beginning and end of words. In addition, there is a short vowel in Kurdish (Row 34 in Table 1) that is not written in the Arabic-based scripting (e.g. in 'دڵ' meaning heart, the phoneme 'i' is between 'د' and 'ڵ') but is written in the Latin-based one.

Although the writing system of Central Kurdish is similar to Persian and Arabic writing systems, it is different from them in a number of letters. The different letters of Kurdish, Persian, and Arabic are given in Table 2. These distinct letters can be used as a feature for language identification as well. The four letters numbered as 41, 42, 43, and 44 in this table are Arabic diacritics that are also used in Persian text but not very commonly. These diacritics cause homograph ambiguity and Kasre problem in Arabic and Persian (Bijankhan *et al.*, 2011), but there are specific corresponding letters for them in Kurdish that appear in the written text (i.e. letters numbered as 31, 32, and 33, respectively). Therefore, this type of ambiguity does not happen in Kurdish.

Table 2 Letters of Kurdish in comparison with Persian and Arabic letters

Language	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	ء	ا	ب	ت	ج	ح	خ	د	ر	ز	س	ش	ع	غ	ف	ق	ک	ل	م	ن	و	ـه
	ā	**A**	**b**	**t**	**j**	**ḥ**	**kh**	**d**	**r**	**z**	**s**	**sh**	**ẹ**	**gh**	**f**	**q**	**k**	**l**	**m**	**n**	**w**	**H**
Kurdish	✓																					
Persian	✓																					
Arabic	✓																					

Language	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
	ی	ژ	پ	چ	گ	ڤ	ڕ	ڵ	ه	ێ	ۆ	ـ	ث	ص	ض	ذ	ظ	ط	◌َ	◌ِ	◌ُ	◌ّ
	Y	**zh**	**p**	**ch**	**g**	**v**	**rr**	**ll**	**a**	**E**	**o**	**i**	**th**	**ṣ**	**ḍ**	**dh**	**ẓ**	**ṭ**	**a**	**e**	**o**	**Shda**
Kurdish		✓				✓							–									
Persian		✓				–							✓									
Arabic		–				–							✓									

3 AsoSoft Text Corpus

In this section, we introduce the first large-scale Kurdish text corpus for Central Kurdish. First, the challenges in building the corpus are presented. Then the steps to building the corpus, including document collection and processing the collected text, are given. Finally, the specifications of the corpus are presented.

3.1 Challenges in building Kurdish text corpus

To collect and process Kurdish textual data for developing a text corpus, several challenges exist that need to be addressed. In this section, we review the most major challenges and present our solutions to them.

3.1.1 *Unavailability of language resources*

The major obstacle to collecting Kurdish text is the unavailability of enough language resources in electronic form. The number of sources to collect Kurdish textual data from is limited in comparison with other languages. Despite this constraint, the unavailability of numerous resources is another challenge. Although there are several publishers and online sources, a majority of them produce or previously used to produce text in non-standard forms such as non-convertible PDF, which cannot be exploited as text files.

The main sources for building a Kurdish text corpus are online Web sites and newspapers, books, and magazines. In building the AsoSoft text corpus, we have used all available sources, including published books and magazines, online Kurdish resources such as online newspapers, Web sites of news agencies, online magazines (including all contents like non-scientific articles, reports, and interviews), all other Web sites with sociological content, analytical articles, and scientific articles. Details of the used sources in our text corpus will be discussed later in this article.

3.1.2 *Crawling complexity*

To collect large-scale data, we needed to crawl all available online Web sites. One of the major problems is the structure of Kurdish Web sites, which is very diverse, making the crawling task a challenge. Due to these variations, it is not straightforward to easily use common crawlers. Therefore, we have been forced to either customize common widely used crawlers such as Apache Nutch and PhpCrawler or to create crawlers for each Web site specifically to get the archives of the Web sites.

3.1.3 *Duplicate crawled data*

Another phenomenon that has been prevented from happening is duplicate text files. This issue used to happen when one single file was crawled more than once due to accessing it through different URL addresses. To handle this issue, filtering repeated data has been done once on the files crawled and also on the corpus altogether. First, the files crawled from each Web site have been checked to filter duplicate pages, which may have been crawled due to accessing them through different URLs. For each crawled document, two substrings whose lengths were not more than the document's length were randomly selected. For documents whose lengths were more than 200 characters, the selected substrings' lengths were not less than 100 characters. Then, we checked any new crawled document against all the selected twin substrings, and in case, a document contained both substrings existing in a twin substring collection, it was identified as a duplicate one and therefore it was filtered. After gathering all the documents, we performed that strategy once again on all the corpus documents.

3.1.4 *Mistaking Arabic and Persian content for Kurdish*

As mentioned in Section 2, the writing system of Central Kurdish is Arabic-based, which is also identical to that of Persian. The similarity of the Kurdish writing system to that of Persian and Arabic results in collecting non-Sorani documents in automatic exploring of sources, especially in online crawling. There is a large number of Web sites containing also non-Kurdish documents (mainly Persian and/or Arabic) without any tags for their distinction. To prevent Arabic and Persian texts from being crawled and saved into the corpus, we first identified the language by using alphabetical differences that Central Kurdish has with Persian and Arabic. As shown in Table 2, Central Kurdish has some letters that are not used in Arabic and Persian, and there are some letters of Arabic and Persian writing systems that do not exist in the Kurdish writing system. Each of these three languages has several language-specific letters that make this method helpful and efficient in the language identification task. Furthermore, the average document length of the corpus is 410 tokens, and the average token length is 7.3 letters, which result in about 3,000 letters per document. Also, the aforementioned exclusive letters form a great portion of the Kurdish alphabet. To ensure correctness of this assumption, we calculated the ratio of the number of Kurdish-specific letters (six letters as shown in Table 2) to the number of all letters for 4,600 documents. This ratio is 16.7%, which means we can detect Kurdish documents with a high level of confidence. On the other hand, even if a Kurdish document does not contain those letters, it probably is too short, in which case, the document loss cost is not very much.

3.1.5 *Processing and conversion of data to a standard format*

Another challenging problem in building the corpus is the diversity of writing forms of the documents. Due to unavailability of a standard reference for Kurdish scripting and also unawareness of some Kurdish authors, it is no exaggeration if we say that every Web site and publisher (and even writers and authors) has their own standard in writing Kurdish text. Another challenge in processing Kurdish text is the misuse of letters and also coding problems for special characters (this problem is similar to that of Persian). These unpleasant mentioned facts are the origins of many processing problems of AsoSoft text corpus, such as normalization. These problems and our solutions to them are addressed in detail in Section 4.

We have processed all the documents gathered from different sources and converted them to a unified structure of text files. The final documents of the corpus are adapted to the AsoSoft text corpus template, which is a Text Encoding Initiative (TEI) XML format (TEI, 2016). This task is described in Section 3.4.

3.2 Document collection

Generally, to build a text corpus for a language, the collected text has to be a representative sample of that language, i.e. the corpus should cover text with a range of different contents having a credible size to be balanced without having any bias. To address these features, we have tried to gather a collection of data from various resources in AsoSoft corpus. The sources include books (academic, political, scientific, historical, linguistic, novels, etc.), magazines (articles, interviews, etc.), and Web sites (news, reports, articles, etc.). However, due to unavailability of electronic textual documents from a number of resources, such as school books and official letters, Web sites form the main source of the corpus. A summary of the main sources of data collection in the AsoSoft text corpus is given in Table 3.

It has to be mentioned that crawled online documents contain a variety of document types such as news articles, interviews, reports, and scientific papers. The crawled Web sites include data from 2003 to 2015. We have crawled fifty-two Web sites that contain Kurdish text. Of course, there are some small Web sites that have Central Kurdish text but have not been crawled. The crawled Web sites cover different types of Web sites with different types of content, including the online version of printed newspapers (e.g. Kurdistany Nwe: knew.org), online news agencies (e.g. emmrro.com), scientific Web sites (e.g. dangakan.info), technological and computer Web sites (e.g. kurditgroup.org), religious Web sites (e.g. dangiislam.com), Web sites with general content (e.g. wikipedia.com), etc. A list of Web sites with the larger size of text is given in Table 4.

3.3 Topic annotation

Defining topic tags for the corpus enriches it to be used in different research studies of language processing applications such as topic identification. A part of the AsoSoft text corpus, including 105,651 documents (~22% of the documents in the corpus), is equipped with this feature. We have used six topic labels of political, social, sports, scientific (technological), literary, and religious. The number of documents and the size of each topic are given in Table 5. As given in the table, some of the tags are of a larger size, for which two reasons may be mentioned. The first reason is a sociolinguistic one special to the Kurdish language, which has resulted in its poor development in certain areas like science; it is too difficult for one to gather, or even to find, a great deal of Kurdish text on scientific subjects either from printed books or online sources. At the same time, Kurdish text on political subjects has been produced much more than any other area, and

Table 3 Sources of the AsoSoft text corpus

Name	No. of tokens (%)	No. of documents	Collected by
Text books and magazines	41 million (21.8)	58,000	Taking from authors and publishers
Web sites	147 million (78.2)	400,000	Crawling the Web sites

Table 4 Larger Web sites used in collecting the AsoSoft text corpus

Rank	Name/Address	No. of tokens	No. of documents	Description
1	Knwe.org	29,959,647	67,220	Newspaper
2	Peyamner.com	19,120,424	110,310	News agency
3	Emmrro.com	14,219,504	14,925	Online newspaper
4	Awene.com	8,271,411	20,231	Newspaper
5	Dangakan.info	7,393,688	7,917	Social articles
6	Penusakan.com	6,860,309	8,759	Cultural/Political
7	Gulanmedia.com	6,068,317	17,597	News agency
8	Kurdstannet.org	5,457,282	5,562	Online newspaper
9	Darunnasi.com	1,507,356	1,808	Scientific
10	Dangiislam.com	1,846,021	1,970	Religious

Table 5 Document topics of the AsoSoft text corpus

No.	Topic	No. of documents	No. of tokens
1	Political	79,417	26,040,705
2	Social	14,925	14,219,504
3	Religious	3,800	2,322,425
4	Sports	13,758	1,838,490
5	Literary	5,174	1,272,523
6	Scientific (technological)	2,007	1,616,585

it definitely is because of special political conditions in which the Kurds live. The second reason is a technical one regarding the sources we have crawled text from; the Web sites mainly lacked a specific structure.

Documents were topically annotated automatically in two ways. If the Web site contained contents of just one topic, the documents crawled from that Web site were tagged with that topic label. Also, if the Web site contained documents of more than one content and the documents had the topic tag, their topic tags were maintained and the documents were tagged accordingly.

3.3.1 *Topic identification*

To assess the topic tag annotation accuracy, we performed a topic identification task on the corpus. To this aim, we selected 1,272,523 tokens from documents of each topic. This number of tokens is chosen to make the data balanced, since the topic tag with the least size of documents, i.e. literary topic, containing the aforementioned number of tokens. We used term-frequency (TF) and term-frequency inverse document-frequency (TF-IDF) (Jurafsky and Martin, 2008) for weighting of the bag of words vector for the 1,000 most frequent tokens extracted from the documents. Then, four machine learning algorithms, including support vector machine (SVM) using the polynomial kernel and sequential minimal optimization (SMO) algorithm, BayesNet, naïve Bayes, and decision tree, were used to train classifiers. In these evaluations, we have used the Weka package (Hall *et al*., 2009).

We used 80% of the annotated data as the training set and 20% of the data as the test set. The performance of the classifiers is calculated using precision, recall, and *F*-measure criteria. The evaluation results for TF are given in Table 6 and for TF-IDF are shown in Table 7. The results show high performances for the methods, indicating the correctness of learning the topics using the data set, which means that the annotation of the topics is done correctly. Also, it needs to be emphasized that the machine learning methods have resulted in high performances, which are probably due to the consistency of labeled data, i.e. the data for each topic are mainly collected from a limited number of sources that have high intraclass uniformity in writing style and vocabulary but have low interclass similarity. The evaluations in this section were repeated using 2.6,000 most frequent tokens as the feature vector from which similar performances were obtained.

3.4 Corpus format

To standardize the data, we have processed and converted the entire collected text from various sources to the TEI format (Dunlop, 1995; TEI, 2016). The processing phase's aim was to correct and normalize

Table 6 Performance evaluation of topic identification using TF

Method	*F*-measure	Recall	Precision
SVM (SMO)	98.2	98.2	98.2
Bayes network	95.7	95.7	95.8
Naïve Bayes	92.7	92.7	92.7
Decision tree	97.1	97.1	97.1

Table 7 Performance evaluation of topic identification using TF-IDF

Method	*F*-measure	Recall	Precision
SVM (SMO)	98.2	98.2	98.2
Bayes network	92.6	92.7	92.7
Naïve Bayes	92.6	92.7	92.7
Decision tree	97.1	97.1	97.1

the text. We faced several challenges in the processing and normalization of the corpus, including the coding of special characters, using non-Kurdish characters, and spelling errors. In Section 4, we have presented the details of the processing steps of the corpus, some of which are novel techniques.

The TEI is an international organization that collectively develops and maintains guidelines for the representation of texts in the digital form. After the processing steps, every document has been converted into the TEI format. In this corpus, only a limited number of tags have been selected to be used; however, using this format enables us to extend the tags in the future in case we need to add extra information to the corpus. A sample file of the corpus in the TEI XML format is given in Fig. 1. A summary of the used tags in this figure and other files of the AsoSoft text corpus is reviewed below.

- asoCorpus: It contains the whole of a TEI-encoded corpus.
- teiHeader: It has been used once for all the corpora containing general information on the corpus and once again for each document. Both uses of this tag for our corpus have been shown in Fig. 1.
- encodingDesc: It shows the relationship between the document and the source it has been collected from.
- editorialDecl: Tags that are used for normalization and encoding of the document are placed in this section.
- normalization: Text normalizations that have been done are cited here.
- revisionDesc: a brief history of processing steps done on the corpus is given here.
- TEI: All the information and tags pertaining to the document are placed here.
- publicationStmt: Tags pertaining to publishing and publisher are placed here.
- Publisher: Organization or institution that has published the document is cited here.
- profileDesc: Non-bibliographic aspects of the document, i.e. information pertaining to the document's topic tag, are presented here.
- textDesc: It includes explanations on the context in which the text is used.
- domain: It includes the document's topic tag.
- date: It refers to the document's publication date.
- title: It refers to the document's title.
- body: The text of the document is placed here.

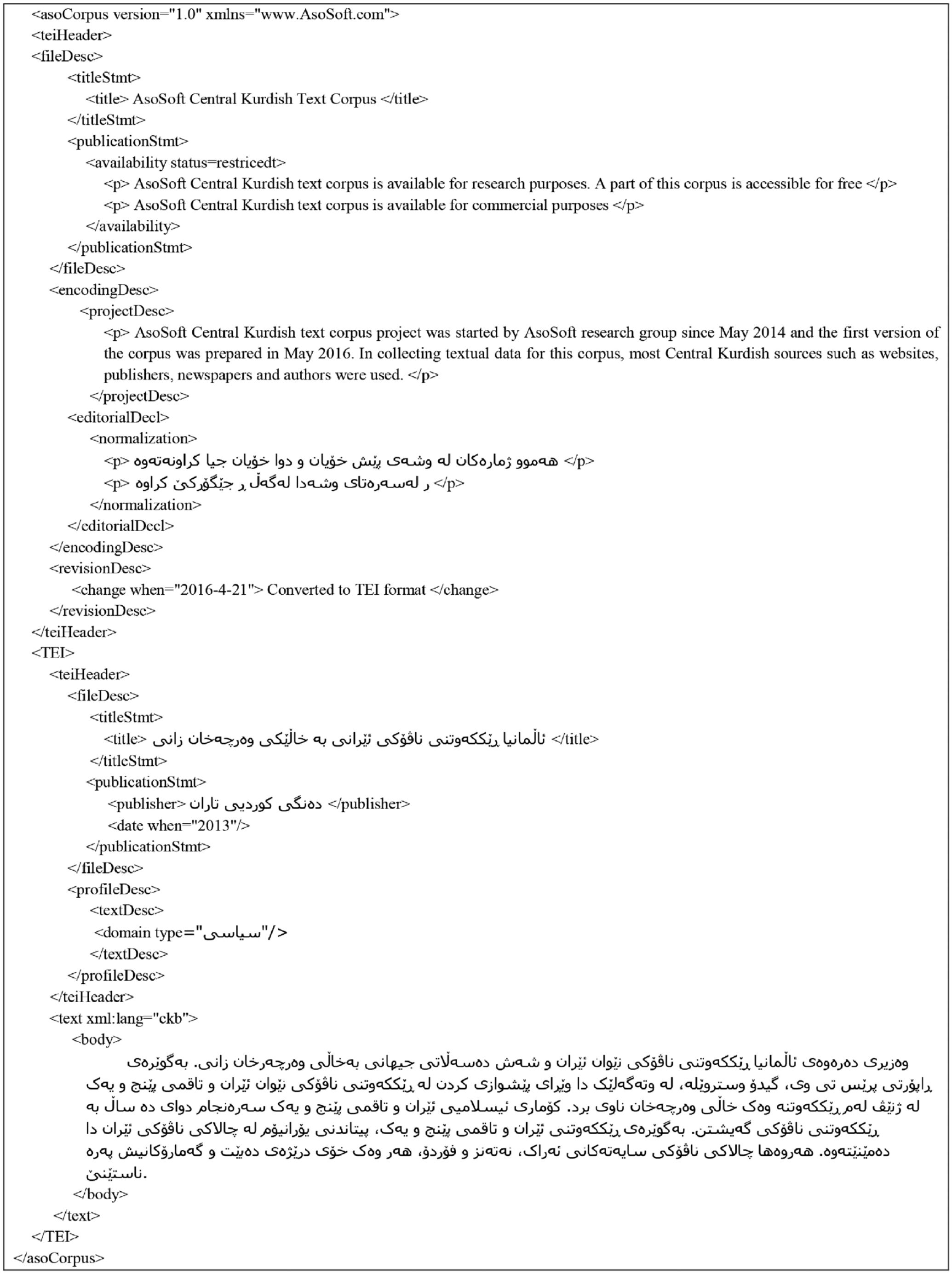
3.5 General specifications of the corpus

After the data collection, we performed the processing steps in Section 4 for correcting and normalizing the corpus. Then, we extracted general information out of the corpus. Tokenization is undoubtedly required for some important statistical information. Tokenization in Kurdish faces several challenges, some of which are drawn upon in Section 4.2.

The general specification of the AsoSoft text corpus after doing the processing steps is given in Table 8. As in the table, the corpus contains about 188 million tokens composed from about 4.66 million unique types. As described in Section 3.3, topic tag for a part of the corpus is annotated. This part contains about 105.7 thousand documents.

Table 8 General specification of the AsoSoft text corpus

Title	Description
Language	Central Kurdish
Corpus size (number of tokens)	188 million
Number of documents	458,000
Average document length (in tokens)	410
Average document length (in characters)	3000
Average sentence length (in tokens)	20.25
Number of unique types	4.66 million
Number of topics	6
Number of documents having topic tag	105.7,000
Corpus format	TEI



```
<asoCorpus version="1.0" xmlns="www.AsoSoft.com">
<teiHeader>
<fileDesc>
      <titleStmt>
         <title> AsoSoft Central Kurdish Text Corpus </title>
      </titleStmt>
      <publicationStmt>
         <availability status=restricedt>
            <p> AsoSoft Central Kurdish text corpus is available for research purposes. A part of this corpus is accessible for free </p>
            <p> AsoSoft Central Kurdish text corpus is available for commercial purposes </p>
         </availability>
      </publicationStmt>
   </fileDesc>
   <encodingDesc>
       <projectDesc>
            <p> AsoSoft Central Kurdish text corpus project was started by AsoSoft research group since May 2014 and the first version of
            the corpus was prepared in May 2016. In collecting textual data for this corpus, most Central Kurdish sources such as websites,
            publishers, newspapers and authors were used. </p>
        </projectDesc>
      <editorialDecl>
          <normalization>
            <p> هەموو ژمارەکان لە وشەی پێش خۆیان و دوا خۆیان جیا کراونەتەوە </p>
            <p> ر لەسەرەتای وشەدا لەگەڵ ڕ جێگۆڕکێ کراوە </p>
          </normalization>
      </editorialDecl>
   </encodingDesc>
   <revisionDesc>
       <change when="2016-4-21"> Converted to TEI format </change>
   </revisionDesc>
</teiHeader>
<TEI>
   <teiHeader>
      <fileDesc>
          <titleStmt>
            <title> ئاڵمانیا ڕێککەوتنی ناڤۆکی ئێرانی بە خاڵێکی وەرچەرخان زانی </title>
          </titleStmt>
         <publicationStmt>
            <publisher> دەنگی کوردیی تاران </publisher>
            <date when="2013"/>
         </publicationStmt>
      </fileDesc>
      <profileDesc>
          <textDesc>
           <domain type="سیاسی"/>
          </textDesc>
      </profileDesc>
   </teiHeader>
   <text xml:lang="ckb">
       <body>
             وەزیری دەرەوەی ئاڵمانیا ڕێککەوتنی ناڤۆکی نێوان ئێران و شەش دەسەڵاتی جیهانی بەخاڵی وەرچەرخان زانی. بەگوێرەی
        ڕاپۆرتی پرێس تی وی، گیدۆ وستروێلە، لە وتەگەلێک دا وێرای پێشوازی کردن لە ڕێککەوتنی ناڤۆکی نێوان ئێران و تاقمی پێنج و یەک
        لە ژنێڤ لەم ڕێککەوتنە وەک خاڵی وەرچەخان ناوی برد. کۆماری ئیسلامیی ئێران و تاقمی پێنج و یەک سەرەنجام دوای دە ساڵ بە
        ڕێککەوتنی ناڤۆکی گەیشتن. بەگوێرەی ڕێککەوتنی ئێران و تاقمی پێنج و یەک، پیتاندنی یۆرانیۆم لە چالاکی ناڤۆکی ئێران دا
        دەمێنێتەوە. هەروەها چالاکی ناڤۆکی سایەتەکانی ئەراک، نەتەنز و فۆردۆ، هەر وەک خۆی درێژەی دەبێت و گەمارۆکانیش پەرە
        ناستێنێ.
       </body>
    </text>
  </TEI>
</asoCorpus>
```

Fig. 1 Sample TEI XML file of the AsoSoft text corpus

After processing and tokenization of the corpus, a list of fifteen most frequent words of Central Kurdish in the corpus was extracted. The list of these tokens, their frequencies, part-of-speech (POS) tag, and meaning in English is given in Table 9.

As can be seen in Table 9, the most frequent tokens are stop words just like other languages. A list of the most frequent non-stop words of Central Kurdish is given in Table 10.

4 Kurdish Text Processing Experiments

In this section, we describe our experiments in the processing of Kurdish text. First, the problems we faced in the processing of the corpus are surveyed and our solutions to them are provided. Second, the tokenization of the corpus and related experimental points are explored. Then, the characteristics of Kurdish language that are important from the computational viewpoint are investigated. Finally, we provide the perplexity analysis of the language according to the corpus and also the Zipf chart is drawn.

4.1 Text processing problems and solutions

There are several challenges in the processing of Kurdish text that affect the tokenization and also research in the field of NLP on these texts. To relax the effects of the related problems, we have performed a number of corrections through processing the corpus and have normalized it. The problems and their solutions are given below. Although some of the

Table 9 Most frequent Kurdish words in the AsoSoft text corpus

Rank	Word	Frequency	POS	Meaning
1	و	8,337,056	Conjunctive	and
2	له	5,505,496	Preposition	from, of, at, in, by, since
3	به	2,892,628	Preposition	with, by
4	که	2,736,939	Conjunctive	the, that, which, when, whom, who
5	بۆ	2,174,376	Noun, preposition, adverb	for, to, why, smell
6	ئەو	1,643,650	Pronoun, determiner	he, she, it, him/her, that
7	ئەم	826,437	Pronoun, determiner	this
8	کوردستان	658,073	Noun	Kurdistan
9	هەر	513,479	Determiner	every, each, any
10	بەڵام	488,828	Conjunctive	but
11	بوو	484,257	Verb	was, became
12	خۆی	463,727	Pronoun	himself/herself
13	لەگەڵ	449,521	Preposition	with, together with
14	لەسەر	448,905	Preposition	on, over, above
15	ئەوەی	446,948	Preposition	that

Table 10 Most frequent non-stop words in the AsoSoft text corpus

Rank in the corpus	Word	Frequency	POS	Meaning
8	کوردستان	658,073	Noun	Kurdistan
26	کورد	325,538	Noun	Kurd
40	ئێران	246,217	Noun	Iran
50	عێراق	212,615	Noun	Iraq
56	سیاسی	190,461	Adjective	political
62	ناو	172,182	Noun, preposition	name, in
63	ساڵی	170,449	Noun	the year
69	هەرێمی	160,458	Noun, adjective	region of, regional
70	سەرۆکی	155,705	Noun	head, president, leader of
72	کەس	152,569	Noun, pronoun	person, nobody

following points are trivial normalization tasks, they are required to be addressed, as they are done for the first time for the Kurdish language. In addition, these problems occur very often in Kurdish text and may cause serious problems in text processing if not solved efficiently. We have also developed a novel algorithm to solve one of the major problems.

4.1.1 *Conversion of non-Unicode Kurdish keyboards*

Non-Unicode Kurdish keyboards still exist and are used by a great number of Kurdish writers, newspapers, etc. In these keyboards, such as in the AliFont keyboard, characters that are exclusive to Kurdish (e.g. ڤ, ڕ, ڵ, ە, ێ, ۆ) are replaced by ASCII characters of Arabic. For example, ۆ is replaced by ؤ. To solve the problem, we used converters for each keyboard/coding to convert text encodings into standard Unicode. The Unicode codes used in the process of conversion are given in Table 1. The difficulty of this conversion is that there are various non-Unicode coding methods used by different document sources that need to be processed separately.

4.1.2 *Multicode Unicode characters*

There is a range of Kurdish Unicode keyboards that use some non-Kurdish Unicode characters. This problem comes from the fact that typography of some letters in Kurdish is the same as those in other Arabic-based writing systems, such as Persian and Arabic writing systems. We replaced common non-Kurdish Unicode characters by Kurdish Unicode characters that look the same. Some of these changes are as follows:

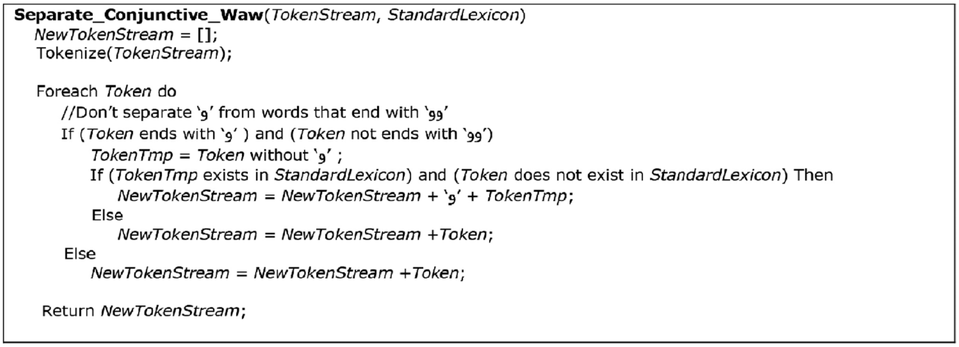
Replacing Arabic ‘ك’ by Kurdish ‘ک’.
Replacing Arabic ‘ي’ by Kurdish ‘ی’.

4.1.3 *Persian silent heh ‘ه’ instead of Kurdish ‘ە’ /a/*

In many words (such as گەورە /gawra/ meaning big), the Kurdish ‘ە’ /a/ is mistakenly written as Persian silent heh ‘ه’ /h/ + zero-width non-joiner (ZWNJ). This problem occurs due to the similar typography of these two letters; however, these letters are different in both character coding (06D5 versus 06BE) and pronunciation (/a/ versus /h/). In addition, Kurdish ‘ە’ is a non-joiner, whereas Persian silent heh ‘ه’ is a joiner. Many Kurdish writers who use non-standard keyboards have to use a Persian silent heh ‘ه’ followed by a ZWNJ (a pseudo-space character) to make Persian silent heh ‘ه’ look like a Kurdish non-joiner ‘ە’. This problem was solved by replacing Persian silent heh ‘ه’ /h/ + ZWNJ by Kurdish ‘ە’ /a/.

4.1.4 *Conjunctive ‘و’ meaning ‘and’ as an independent word*

According to Kurdish linguistic rules, the conjunctive ‘و’ is accounted for as an independent word. However, some Kurdish publishers and scholars are not aware of this fact, and therefore, they join conjunctive ‘و’ to a proceeding word, usually one that ends in a vowel; for instance, in ‘ئێرەو ئەوێ’ (meaning here and there) and ‘چاکەو خراپە’ (meaning good and evil), the conjunctive ‘و’ follows the previous word without a spacing; their correct forms are ‘ئێرە و ئەوێ’ and ‘چاکە و خراپە’. Another wrong form used is to join conjunctive ‘و’ to a proceeding word followed by ZWNJ; this wrong form has occurred 1.2 million times in the corpus. The latter type



```
Separate_Conjunctive_Waw(TokenStream, StandardLexicon)
  NewTokenStream = [];
  Tokenize(TokenStream);

  Foreach Token do
      //Don't separate 'و' from words that end with 'وو'
      If (Token ends with 'و' ) and (Token not ends with 'وو')
          TokenTmp = Token without 'و' ;
          If (TokenTmp exists in StandardLexicon) and (Token does not exist in StandardLexicon) Then
              NewTokenStream = NewTokenStream + 'و' + TokenTmp;
          Else
              NewTokenStream = NewTokenStream +Token;
      Else
          NewTokenStream = NewTokenStream +Token;

  Return NewTokenStream;
```

Fig. 2 Algorithm for correcting the problem of conjunctive ‘و’

could easily be corrected by replacing the ZWNJ character with a space. But to correct the first wrong form, we developed a novel algorithm given below (see Fig. 2) and evaluated its performance. In this algorithm, we needed a standard lexicon; therefore, a lexicon of about 470,000 tokens was taken from the Hanbanaborina dictionary (Hajar, 1989), *Hawler* newspaper (Hawler Press, 2015), and a number of books that we are assured of regarding the correctness of such words as mentioned above in their text. To evaluate the algorithm's performance, 1,000 instances of the corrections made using the algorithm were investigated. Out of 1,000 randomly chosen corrections, 996 instances were correctly modified, which shows a precision of 99.6%.

4.1.5 *Letter 'ر' /ɾ/ as 'ڕ' /r/*

As given in Table 1, there are two types of /r/ in Kurdish with different typographies and different pronunciations. In Central Kurdish, the letter 'ر' /ɾ/ never comes as the first letter of words (Kurdish Academy, 2009), but many Kurdish scholars use 'ر' /ɾ/ instead of 'ڕ' /r/ at the beginning of words, incorrectly, as in 'رێگە' (meaning way) and 'رێککەوتن' (meaning agreement). Therefore, all occurrences of 'ر' /ɾ/ at the start of words were replaced by 'ڕ' /r/.

4.1.6 *The vowel 'وو' /u/ as the first letter of words*

Just like the rule mentioned for 'ر' /r/ previously, the vowel 'وو' /u/ never occurs as the first letter of words according to Kurdish linguistic rules (Kurdish Academy, 2009; Jamal *et al.*, 2013). Therefore, it has, instead, been replaced by the consonant 'و' /w/ at the start of words. For example, the word 'ووشە' (meaning word) in the corpus is corrected as 'وشە'.

4.1.7 *Single 'ـیـ' /i/ instead of its double form 'ـییـ'*

In some words with a doubled 'ـیـ', (i.e. 'ـییـ'), the word has been typed wrongly having a single 'ـیـ'. A common example of this problem is the word 'نییە' (the negative past tense of 'to be'), which is written as 'نیە' in many cases. In the corpus, the case 'نیە' has been corrected as 'نییە'. These two forms of the word 'نییە' do not differentiate in the absence versus presence of one 'ـیـ'; therefore, we can make them uniform.

4.1.8 *Using underline for stretching words*

In some cases, underline character is used wrongly to stretch words, for example 'بیر' (meaning thought) is written wrongly as 'بـــــیر'. We have identified and removed all such characters in the words.

4.1.9 *Punctuation marks*

Punctuation marks have mainly had two problems. First, some of them have been typed in their English forms, like English question mark '?', which should be typed as '؟' in Kurdish. Second, punctuation marks have, in many cases, been incorrectly isolated from their proceeding words; punctuation marks are supposed to be typed immediately after their proceeding word and be followed by a space after them. These two problems have been corrected in the corpus as well.

4.1.10 *Special characters and non-textual symbols*

Since data were collected from different sources, mainly from Web sites and books, these contained many non-textual symbols, special characters (such as those used in Microsoft Word and other text editors for formatting), URLs, and e-mail addresses. All of these symbols and characters have been identified and removed. In removing patterned parts such as URLs and e-mail addresses, we have used regular expressions to identify the patterns and remove them.

4.1.11 *Sticking non-Kurdish letters/digits together with Kurdish letters*

In some cases, digits and non-Kurdish letters (such as English letters) are concatenated to Kurdish words without spacing. These cases have been identified and separated by using regular expressions.

A summary of the corrected mistakes in the corpus with their frequencies is given in Table 11.

4.2 Tokenization

The act of segmenting text to the morphemes included in it is called tokenization (Jurafsky and Martin, 2008). Tokenization as a primary step in text processing is mainly performed by using the space character as a delimiter, although different languages may have different and various forms of proper tokenization. Tokenization in Kurdish faces some challenges, which are mainly due to its complex morphology and its Arabic-based writing system. Kurdish morphology, on the one hand, allows making compound words through combining simple words, which results in multiunit tokens (MUTs), and, on the other hand,

Table 11 Summary of corrections and their frequencies in the AsoSoft text corpus

Wrong form	Corrected form	Frequency
Arabic 'ك'	Kurdish 'ک'	23,006,647
Arabic 'ي'	Kurdish 'ی'	754,161
Persian 'ه'(h)+ZWNJ	Kurdish 'ە' (a)	11,367,642
Conjunctive 'و' joined to a proceeding word	Word + 'و'	1,612.314
Letter 'ر' (r) at the beginning of a word	'ڕ' (rr)	268,805
Vowel 'وو' (u:) as the first letter of words	'و' (w)	73,417
Negative past tense of 'to be', 'نیە'	'نییە'	210,542
Use of 'underline' for stretching words	Deleted	695,739
Punctuation marks: non-Kurdish forms + incorrect spacing	Converted to Kurdish form + corrected the spacing	1,577,961
Special characters and non-textual symbols in text	Removed non-textual symbols and characters	+10,000,000
Concatenation of non-Kurdish letters/digits to Kurdish letters	Separated non-Kurdish letters/digits from Kurdish letters	+10,000,000

allows the abundant use of affixes. With that in mind, we find Sorani's writing system a bit incapable of satisfying the needs of such complex morphology in terms of writing an easily readable text. Besides, there is no standard lexicon to be used as the reference for tokenization in Kurdish. Therefore, we required several considerations in the tokenization of the AsoSoft text corpus, in addition to the corrections mentioned before. In the following, we address the most important considerations in the tokenization of Kurdish text.

- Similar to such other languages as Persian, a single compound word could be written in various ways, which are a result of the concatenative and non-concatenative writing of MUTs. For instance, the infinitive 'دەستنیشانکردن' /dast+niʃɑn+kɪrdɪn/ (meaning determining) can also be written in the following MUT forms: 'دەست نیشان کردن', 'دەستنیشان کردن', and 'دەست نیشانکردن', the difference among all of which is the spacing. This problem, that space is not a reliable delimiter, is even more evident in writing derivative words due to the abundance of affixes in Kurdish, for instance in 'بێ سنوور', 'بێسنوور', 'بێسنوور' / besɪnur/ (meaning borderless, limitless).
- Another kind of compound word in Kurdish is formed through repetition of a simple word, such as 'پۆل پۆل' /pol pol/ (meaning in groups), 'چین چین' /tʃin tʃin/ (meaning layer by layer), etc., which are instances of reduplication. There are different ways of writing this kind of compound word. We accounted for them as a single word in tokenization. Similarly, those compound words that have been formed using the character '-' have been accounted for as one token, such as 'دەستە-دەستە'/dasta-dasta/ (meaning group by group).
- We also considered a number of punctuation marks, including [,: ; ? !], as delimiters and ignored other punctuation marks, including ["" () [] {}], for tokenization. We have been careful of using 'dot' as a delimiter because in some cases such as abbreviations (e.g. 'ه.ش' meaning Hijri-Shamsi), this punctuation mark is used as a part of the token. Therefore, we listed the most common abbreviations used in Kurdish text not to consider the dot as a delimiter wrongly. In addition, we were aware of the dot in digits that contained decimal points.
- For MUTs, some writers often prefer to use a ZWNJ (i.e. pseudo-space) character between the tokens, similar to what is done in Persian; however, using ZWNJ is not recommended in Kurdish (DepIT-KRG, 2015). Furthermore, there are other forms of MUTs in the text. For example, the words 'لە کاتێکدا' /la kɑtekdɑ/ (meaning while) and 'دەستنیشانکردن' have various forms and frequencies in the corpus after doing the normalizations as shown in Table 12. In this table, '^' means ZWNJ and '_' means space character. The reason that the frequencies of the last three cases for 'لە کاتێکدا' (and also the last two cases for 'دەستنیشانکردن') are zero is the fact that we have replaced all Persian silent heh 'ه' + ZWNJ by Kurdish 'ە' /a/ in the corpus, as mentioned in Section 4.1.

Table 12 Orthographic variations of MUTs ‘لە کاتێکدا’ and ‘دەستنیشانکردن’ with their frequencies in the AsoSoft text corpus after normalization

Written form	Frequency	Written form	Frequency
لە_کاتێکدا	24,747	دەستنیشانکردن	5,094
لەکاتێکدا	32,749	دەستنیشان_کردن	1,148
لە_کاتێک_دا	2118	دەست_نیشان_کردن	621
لەکاتێک_دا	423	دەست_نیشانکردن	518
لە_کاتێک^دا	256	دەست^نیشان_کردن	7
لەکاتێک^دا	51	دەست^نیشان^کردن	5
لە^کاتێک^دا	0	دەست_نیشان^کردن	2
لە^کاتێک_دا	0	دە^ستنیشانکردن	0
لە^کاتێکدا	0	دە^ستنیشان_کردن	0

Also, it has to be clarified that to extract the statistics reported in this article, we have replaced all ZWNJ characters by space character for consistency.

4.3 Computational characteristics of Kurdish

In addition to the aforementioned points on the processing of Kurdish, other characteristics of this language are hereby summarized. These features have been used implicitly or explicitly during processing the corpus and can be used in other NLP applications as well.

4.3.1 *Phonographic system*

An interesting general property of Kurdish language (and also the Central Kurdish) is the correspondence of each written letter to the spoken phoneme. This feature helps us to generate a computational lexicon automatically in applications like speech recognition and also to routinely pronounce text in a text-to-speech system. The only exception of this property in Kurdish is the /i/ phoneme that is also called bzroka (i.e. the last letter in Table 1). This phoneme is a very short vowel that, like all other vowels in Kurdish, occurs between two consonants and is not written in the Arabic-based writing system. The duration of this vowel is so short that some linguists ignore it.

4.3.2 *Homographs*

Like other languages, Kurdish also includes homograph words that have different meanings for a same written form. However, since Kurdish is a phonographic language, homographs cannot be heteronyms (i.e. having different pronunciations and different meanings) but can be homonyms (i.e. having the same pronunciation and different meanings). Examples are given in Table 13.

4.4 Perplexity of Central Kurdish

Perplexity is a well-known intrinsic evaluation metric for statistical language models (LMs), which defines the weighted average branching factor of a language (Jurafsky and Martin, 2008). The branching factor of a language is defined as the number of possible next words that can follow any given word. In this article, we constructed N-gram LMs of Kurdish language from our corpus and then calculated the perplexity of this language. Based on the best of our knowledge, it is the first time that perplexity is calculated for Kurdish. The value of this metric gives us an estimation of the complexity of Kurdish especially in comparison with other languages.

To calculate the perplexity, standard N-gram LMs are computed. This statistical LM assigns a probability to every word in lexicon using the sequences of words appearing in the corpus. In the calculation of the probability values, it is assumed that there is no dependency between consecutive words in sentences (for unigram, $N = 1$) or a word is only depending on one previous word (for bigram, $N = 2$) or depending on the two previous words (for trigram, $N = 3$). Such LMs are essential components of many NLP applications such as speech recognition or machine translation (Huang *et al.*, 2001).

We randomly selected a number of sentences including 10% of the corpus (about 18 million tokens) as the test set after normalization steps were done. The remaining 90% of the corpus (about 170 million tokens) was used as the training set. A lexicon was extracted from the corpus for the experiments that included 100,000 of the most frequent tokens. The frequency of the lexicon tokens was not less than ninety-seven. The standard N-gram LMs (for $N = 1$, 2, and 3) were estimated using the training set by maximum likelihood estimation and smoothed with Witten-Bell discounting. To build the LMs described in this article, the CMU statistical language modeling toolkit (Clarkson and Rosenfeld, 1997) was used. In these experiments, the cutoff value for both bigram and trigram was two. The calculated perplexities of

Table 13 Examples of Kurdish homographs and their frequencies in the AsoSoft text corpus

Word	Pronunciation	Meaning 1	Meaning 2	Meaning 3	Frequency
بۆ	/bo/	Why	For	Smell	2,411,135
سەر	/sar/	Head	On		384,355
ناو	/nɑw/	Name	Inside		172,182
گرتن	/gɪrtɪn/	Hypothesize	Jail	Get	10,777
پێشکەوتن	/peʃkawtɪn/	Moving ahead	Progress	Success	7,702
دیاردە	/dɪyɑrda/	Phenomenon	Visible		5,308
بار	/bɑr/	Load	State		5,989
پار	/pɑr/	Last year	Section		4,304
تاڵ	/tɑɫ/	Bitter	Cloudy weather	Thread	3,732
کوڵ	/kʊɫ/	Short	Boiling		531

Kurdish on the test set for unigram, bigram, and trigram are given in Table 14. In this table, the perplexity and also entropy values are reported for two cases: in the first one, the out-of-vocabulary (OOV) tokens were ignored (shown as NoOOV in the table and are given between parentheses), and in the second case, all OOV tokens were considered as an *unknown* word and probability was calculated for it. The entropy, as another evaluation metric for LMs, was calculated as the logarithm of the perplexity. As expected, the perplexity values of the NoOOV case are higher than the other case. Also, the perplexity is reduced by increasing the order of N-gram model that shows that the trigram model is a better LM than bigram and monogram. A better LM predicts the details of the test data better.

Table 14 Perplexity of Central Kurdish language using N-gram language models for N = 1, 2, 3

	Unigram (NoOOV)	Bigram (NoOOV)	Trigram (NoOOV)
Perplexity	13,209.8 (29,101.7)	648.3 (994.9)	276.0 (407.8)
Entropy	13.7 (14.8)	9.3 (9.9)	7.1 (8.7)

To compare the complexity of Kurdish with other languages based on the perplexity metric, the perplexities of the trigram model for this language and two other languages, English and Persian, are given in Table 15. The perplexity of English has been calculated using a 20,000 lexicon, a training set of 38 million tokens, and a test set of 1.5 million tokens from Wall Street Journal (WSJ) data (Jurafsky and Martin, 2008). For Persian, training and test sets consist of 100 million and 10,000 tokens, respectively, and a small lexicon of size 1,000 tokens has been used (Sameti et al., 2011). Despite the fact that the evaluation conditions for the three languages in Table 15 are not the same, the perplexity values indicate only a general comparison of these languages. In English, WSJ is a uniform corpus constructed from purely journalistic data and Persian corpus is a multidomain corpus collected from various sources (like our corpus). The higher perplexity of Kurdish probably comes from the fact that this language has a complex morphological structure (Hosseini *et al.,* 2015). The morphological structure of Kurdish language allows more complicated word-forming than the English and even the formal Persian language (Naserzade, 2018). The informal Persian structure also permits to generate a multipart token, which is more similar to the Central Kurdish language structure.

4.5 Zipf's law for Central Kurdish

According to Zipf, human beings favor less effort for more advantages; this shows that people want to be economic, and this tendency is obvious in different levels of language production from phonology to lexicon. Therefore, based on this law, both the speaker and the hearer tend to face less effort in their linguistic communications; this needs them to have a few number of more frequent words, a middling number of medium-frequency words, and a large number of less frequent words (Manning and Schütze, 1999).

Based on Zipf's law, there is a relationship between word frequencies and their ranks in the lexicon; i.e. , where *f* stands for word frequency and *r* is word's rank in the list of the words. Commonly, Zipf's law indicates that for a natural language, the most frequent word occurs approximately twice as often as the second most frequent word, three times as often as the

Table 15 Perplexity of Central Kurdish language in comparison with English and Persian

Language	Evaluation condition			Perplexity
	Train set	**Test set**	**Lexicon**	**Trigram**
Sorani Kurdish	170 million	18 million	100,000	276
Persian	100 million	10,000	1,000	135
English	38 million	1.5 million	20,000	109

third most frequent word, and etc. We evaluated the mentioned law on our text corpus empirically to see if the relationship between word frequency and rank holds. As shown in Table 9, the word 'و' is the most frequent word (8,337,056 occurrences out of 188 million tokens, about 4.4% of all words), and by Zipf's law, the second-place word will be approximately half of accounts of the first-place word, while the real count is slightly different (5,505,496 occurrences, nearly 2.9%), and the third-place word nearly follows the Zipf's law (2,892,628 occurrences, about 1.5%). Table 9 shows that the fourth-place and fifth-place words do not follow the Zipf rule, probably due to the normalization problems.

Figure 3 shows the word–frequency relationship in our corpus on a log-log plot. According to Zipf's law prediction, such a graph should be in the form of a straight line with slope −1. As can be seen, the line in this figure roughly fits a straight line with slope −1.

5 Summary and Future Works

In this article, we introduced the first large-scale Kurdish (Central Kurdish) text corpus, AsoSoft text corpus. Also, the experiments and lessons from building this corpus were presented. We met several challenges in collecting and processing the text corpus and reported our solutions to them. A part of the corpus was topically annotated for special NLP applications such as document classification. The standard N-gram LMs were constructed from the normalized text, and the perplexity of Central Kurdish was calculated and

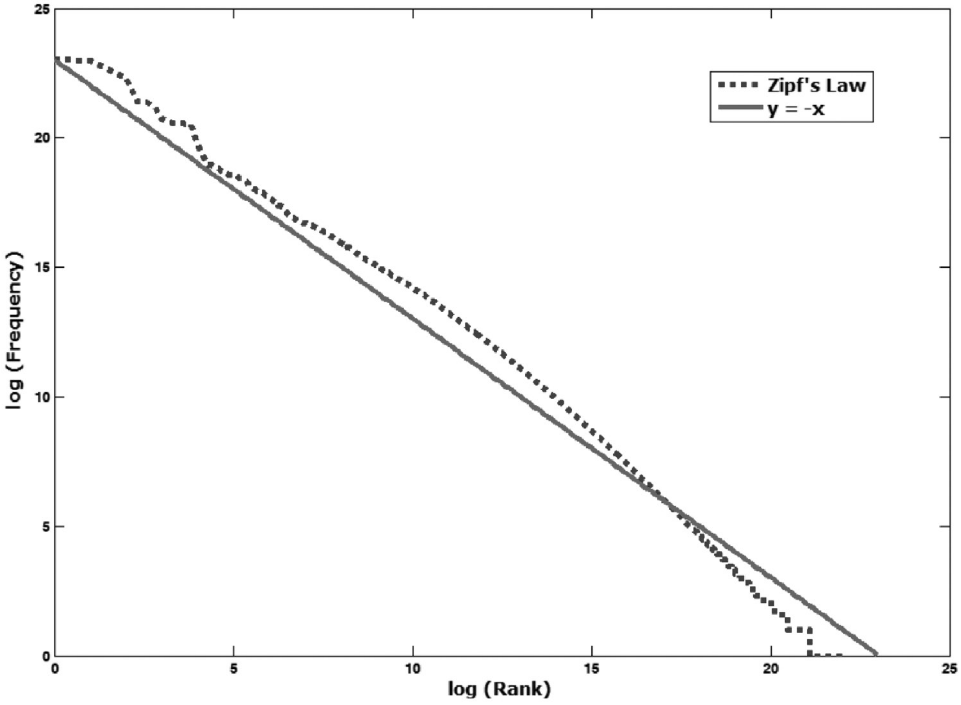


Fig. 3 Zipf's law evaluated on the AsoSoft corpus for Central Kurdish

compared with the perplexity of English and Persian languages. Also, Zipf's distribution of Central Kurdish was presented. Although several corrections and normalizations were performed on the corpus, more normalization is required to correct misspelled and non-standard tokens. In addition to doing more refinements on the corpus, we are collecting more text to enrich the corpus. We believe that the AsoSoft text corpus, as the first large-scale computational linguistics resource of Central Kurdish, will encourage the researchers and businesses to work on this language.

Acknowledgements

The authors would like to thank all the people and organizations who helped them in collecting the corpus by providing their documents, among them are Mukryani Publisher, Jin Institute, Kurdish Heritage Institute, and Sardam Publisher.

References

Barkhoda, W., **ZahirAzami, B.**, **Bahrampour, A.**, **and Shahryari, O. M.** (2009). A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Ajman, UAE, pp. 557–562.

Sheykh Esmaili K. and Salavati Sh. (2013). Sorani Kurdish versus Kurmanji Kurdish: An empirical comparison. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria. 300-305.

Sheykh Esmaili K., Salavati Sh., Yosefi S., Eliassi D., Aliabadi P., Hakimi Sh., and Mohammadi A. (2013). Building a Test Collection for Sorani Kurdish. In Proceedings of the 10th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA'13). Fez, Morocco.

TEI (2016), TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.0.0 , Text Encoding Initiative Consortium.

Bijankhan, M., **Sheykhzadegan, J.**, **Bahrani, M.**, **and Ghayoomi, M.** (2011). Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, **45**: 143–164.

Clarkson, P. and Rosenfeld, R. (1997). Statistical language modeling. In *Proceedings of EuroSpeech'97*, Rhodes, Greece, pp. 2707–10.

Dep. of IT, Kurdistan Region Government (KRG). (2015). http://unicode.ekrg.org/ku_unicodes.html (accessed July 2015).

Dunlop, D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computer and the Humanities*, **29**: 85–98.

Gautier, G. (1998). Building a Kurdish language corpus: an overview of the technical problems. In *Proceedings of ICEMCO 98 6th international conference and exhibition on multilingual computing, Cambridge, UK*.

Geoffrey, H. (2002). The corpus of contemporary Kurdish newspaper texts (CCKNT): a pilot project in corpus linguistics for Kurdish. *Kurdische Studien*, **1**: 148–155.

Hajar. (1989). *Hanbana Borina Kurdish-Persian Dictionary*. Iran: Sorush Publication.

AQ16

Hall, **M.**, **Frank**, **E.**, **Holmes**, **G.**, **Pfahringer**, **B.**, **Reutemann**, **P.**, **and Witten**, **I. H.** (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, **11**: 10–18.

Hawler Press. (2015). http://www.hawler.in.

Hosseini, H., **Veisi, H.**, **and MohammadAmini, M.** (2015), *KSLexicon: Kurdish-Sorani Generative Lexicon, The First National Conference on Corpus-based Linguistics* (in Persian), Tehran, Iran, pp. 33–50.

Huang, X., **Acero, A.**, **and Hon, H. W.** (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River: Prentice Hall PTR.

Ilkhanizadeh, M. (2006). *Learn to Read and Write Kurdish Language* (in Persian). Iran: Kurdistan Publication.

Jamal, A., **Woria, O. A.**, **Farouq, O. S.**, **and Azad, A. M.** (2013). *Kurdish Orthography Based on Scientific Basis of Public Orthography* (in Kurdish). Hawler, Kurdistan: The Kurdish Academy.

Jurafsky, **D. and Martin**, **J. H**. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edn. Prentice Hall.

Kaveh, M. (2005). *Kurdish Linguistic and Grammar* (Saghezi *Accent)* (in Persian). Iran: Ehsan Publication.

Kurdish Academy. (2009). *Recommendations of 'Toward a Unified Orthography' Conference* (in Kurdish). Hawlere, Kurdistan: The Kurdish Academy.

Manning, **C. D. and Schütze**, **H.** (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Naserzade, **M.** (2018). *Designing and Implementation of a Morphological Analyzer for Central Kurdish Language* (in Persian). M.S. thesis, Sharif University of Technology, Tehran-Iran.

Nebez, J. (1993). The Kurdish Language from Oral Tradition to Written Language. http://www.kurdishacademy.org/?q=node/135.

Peyamner. (2015). Peyamner News Agency. www.peyamner.com.

Rokhzadi, A. (2011). *Kurdish Phonetics and Phonology* (in Persian). Tarfand Press.

Ruwange Blog. (2015). http://ruwange.blogspot.com.

Sameti, **H.**, **Veisi**, **H.**, **Bahrani**, **M.**, **Babaali**, **B.**, **and Hosseinzadeh**, **K.** (2011). A large vocabulary continuous speech recognition system for Persian language. *Springer EURASIP Journal on Audio, Speech, and Music Processing*, **2011**(1), 1–12.

Sheykh Esmaili, K. and Salavati, Sh. (2013). Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, pp. 300–305.

Sheykh Esmaili, K., **Salavati, Sh.**, **Yosefi, S.**, **Eliassi, D.**, **Aliabadi, P.**, **Hakimi, Sh.**, **and Mohammadi, A.** (2013). Building a test collection for Sorani Kurdish. In *Proceedings of the 10th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA'13)*. Fez, Morocco.

TEI. (2016). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Version 3.0.0, Text Encoding Initiative Consortium.

VOA. (2015). Voice of America - Kurdish (Sorani). www.dengiamerika.com.

Walther, G. (2011). Fitting into morphological structure: accounting for Sorani Kurdish endoclitics. In M. Stefan (ed.), *The Proceedings of the Eighth Mediterranean Morphology Meeting (MMM8)*, Cagliari, Italy, pp. 299–322.

Walther G. and Sagot B. (2010). Developing a large-scale lexicon for a less-resourced language. In *SaLTMiL's Workshop on Lessresourced Languages (LREC)*.

Walther, **G.**, **Sagot**, **B.**, **and 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*. Belgrade, Serbia.

Wynne, M. (ed.) (2005). *Developing Linguistic Corpora: A Guide to Good Practice*, vol. **92**. Oxford: Oxbow Books.

Note

1. A part of this corpus is available for researchers from the AsoSoft group's Web site, www.asosoft.com.