A Tokenization System for the Kurdish Language

Sina Ahmadi
Insight Centre for Data Analytics
National University of Ireland Galway, Ireland
ahmadi.sina@outlook.com

Abstract

Tokenization is one of the essential and fundamental tasks in natural language processing. Despite the recent advances in applying unsupervised statistical methods for this task, every language with its writing system and orthography represents specific challenges that should be addressed individually. In this paper, as a preliminary study of its kind, we propose an approach for the tokenization of the Sorani and Kurmanji dialects of Kurdish using a lexicon and a morphological analyzer. We demonstrate how the morphological complexity of the language along with the lack of a unified orthography can be efficiently addressed in tokenization. We also develop an annotated dataset for which our approach outperforms the performance of unsupervised methods[1].

1 Introduction

A text, as the input of text processing applications, is composed of a string of characters and is interpreted based on the way it is segmented. Words and sentences are two segments in a text which carry meaning at different levels. Although the boundaries of words and sentences are specified to some extent in some scripts, e.g. by using whitespaces and punctuation marks, finding such boundaries is a non-trivial task (Guo, 1997). For instance, in scripts where spaces are not widely used or the *scriptio continua* style is used, such as Japanese, Chinese and Classical Latin, or languages where words are concatenated to create compound forms as in German, word boundary may not be explicitly specified.

In natural language processing (NLP), tokenization generally refers to the task of finding segment boundaries in a text. More specifically, retrieving the boundary of words and sentences are respectively known as word tokenization and sentence tokenization. Given a string of characters, a tokenization system, also known as lexical analyzer or tokenizer, splits the input into tokens, i.e. words or sentences (Kaplan, 2005). Tokenization is one of the most important and fundamental language processing tasks with many applications, such as part-of-speech tagging and machine translation (Webster and Kit, 1992).

Given the recent advances in NLP and artificial intelligence, tokenization is considered a solved problem and has been efficiently addressed for many languages (Habert et al., 1998; Forst and Kaplan, 2006). Although methodologies and approaches in tokenization of one language might be applicable to and beneficial for another language, linguistic and orthographic issues can make tokenization a language-specific problem (Lebart et al., 1997). For instance, although whitespaces are generally used in Arabic-based scripts, such as Urdu, Persian and Arabic, the joining and non-joining characteristics of graphemes create further complexity in tokenizing compound words, i.e. words consisted of more than one word, and various morphemes, such as affixes and clitics (Rehman et al., 2013; Shamsfard et al., 2009).

In this paper, we carry out a preliminary study on the task of tokenization for the Kurdish language with a particular focus on two of the Kurdish dialects, i.e. Kurmanji and Sorani, for which Latin and Arabic-based scripts are respectively used. We show how the Kurdish scripts and their lack of standardized orthographies create variations in writing words, especially compound forms. To address this task, we develop a tokenization system using a basic morphological analyzer and a lexicon and demonstrate that it outperforms regular expression based and unsupervised neural methods.

[1] The tool and the resources are publicly available at https://github.com/sinaahmadi/KurdishTokenization

2 Related Work

The notion of word is one of the most basic concepts in various fields and therefore can be defined in different ways. Generally speaking, a word refers to a building block of a sentence. However, from a morphological point of view, a word, which is also known as word-token, is defined based on its form and meaning. If a word carries a concrete meaning, it is defined as a word-form such as *drives*, *driving*, *drove*, while a word with an abstract meaning is known as a lexeme or lexical item, e.g. DRIVE (Haspelmath and Sims, 2013, p 15). Lexemes are also distinguished by their function as headwords in dictionaries. It is worth mentioning that in addition to lexemes, lemmas are used to refer to the canonical forms of the lexemes. For instance, although خواردن (*xwardin*) EAT and خواردنەوە (*xwardinewe*) DRINK are two distinct lexemes in Kurdish, they both have one lemma and that is *xwardin*. The task of retrieving word lemmata is called lemmatization and is of importance in NLP.

Analogous to the distinction between word-forms and lexemes in morphology, corpus linguistics distinguishes a word as token and type based on their distinctness in a text. While a token can frequently occur, a type is considered the unique form of the token which can also be used as a dictionary entry (McEnery and Wilson, 2003). Additionally, Habert et al. (1998) describe tokens based on the lexicographic information, the context such as sub-languages and terminologies and, the applications. The application-based definition suggests that word boundary depends on the underlying application for which the tokenization task is required. For instance, the performance of tokenization methods have been studied in various tasks, such as statistical and neural machine translation (Zalmout and Habash, 2017; Domingo et al., 2018), text classification (Hiraoka et al., 2019) and named-entity recognition (Bai et al., 2001).

Being a basic task in information retrieval, text processing and NLP, the task of tokenization has been widely previously studied for many languages. A wide range of techniques are used for the task, particularly rule-based (Marcus et al., 1993; Dridan and Oepen, 2012; Waldron et al., 2006), statistical (Kiss and Strunk, 2006; McNamee and Mayfield, 2004) and more recently, neural networks (Kudo and Richardson, 2018; Schweter and Ahmed, 2019). The latter are particularly beneficial to alleviate open vocabulary problems independent of the language. Moreover, given the importance of tokenization in downstream applications, tokenization tools are usually also provided within NLP frameworks such as Moses (Koehn et al., 2007) and OpenNMT (Klein et al., 2017) for machine translation.

As the earliest work that addresses tokenization for Kurdish language, Rezaie (2001) discusses some of the issues in word boundary in the Arabic-based scripts. Although Kurdish tokenization has been partially addressed in the context of other tasks, such as text classification (Rashid et al., 2017), machine translation (Forcada et al., 2011) and syntactic analysis (Gökırmak and Tyers, 2017), no previous study is found to explicitly focus on Kurdish tokenization. For Kurdish as a less-resourced language, we believe that the current study will pave the way for further developments in Kurdish language processing.

3 An Overview of the Kurdish Language

Kurdish is a less-resourced Indo-European language spoken by 20-30 million speakers in the Kurdish regions of Iran, Iraq, Turkey and Syria (Ahmadi et al., 2019). There are various points of view regarding the classification of the dialects of Kurdish (Matras, 2017). However, Northern Kurdish, also known as Kurmanji, Central Kurdish, also known as Sorani and, Southern Kurdish are less controversially accepted as the Kurdish dialects (Matras, 2019). Historically, many alphabets have been used for writing Kurdish among which the Latin-based and Arabic-based scripts are still widely in use (Chyet and Schwartz, 2003). Although the standardization of the language, in the written and spoken forms, have been a matter of discussion in academia and also among Kurdish people, there is no consensus regarding what is meant by a standard writing system or orthography for Kurdish (Tavadze, 2019). Due to the lack of standardization, different scripts may be used for writing in various dialects. Regarding the popularity of the scripts, the Arabic-based alphabet is widely used for Sorani and Southern Kurdish while the Latin-based is used for Kurmanji. Due to Southern Kurdish being an under-documented dialect (Fattah, 2000), we only focus on the Sorani and Kurmanji dialects in this study.

Table 1 provides a comparison between the Latin-based and Arabic-based alphabets of Sorani and Kurmanji Kurdish. It should be noted that the vowel *i* does not have an equivalent grapheme in the Arabic-

based alphabet. On the other hand, some of the consonants in the Latin-based alphabet are composed of a punctuation mark, usually an apostrophe as in *'e*. In addition to the consonants and vowels, some of the punctuation marks in the two alphabets are provided in Table 1c. Variations are specified by "/".

Similar to the Arabic-based scripts of Persian and Urdu, the Arabic-based script of Kurdish has a zero width non joiner (ZWNJ, U+200C) character which enables joining characters be written in their non-joining grapheme. For instance, the character ڵ in هەڵگرتن (*heɫgirtin*) 'to lift' appears as هەڵ گرتن when followed by a ZWNJ. Moreover, to further add to the length of the joining between graphemes, a dual-joining grapheme known as *Tatweel* or *Kashida* (U+0640) is used. This grapheme does not represent any phoneme but only elongates characters for justification and alignment of the text.

Latin	b	ç	c	d	f	g	h	j	k	l	ɫ/ll	m	n	p	q	r	ř/rr	s	ş	t	v	w	x	y	z	‘/’e/ë	ẖ/‘h	ẍ/x	’
Arabic	ب	چ	ج	د	ف	گ	ه	ژ	ک	ل	ڵ	م	ن	پ	ق	ر	ڕ	س	ش	ت	ڤ	و	خ	ی	ز	ع	ح	غ	ئ

(a) consonants

Latin	a	e	ê	i	î	o	û	u
Arabic	ا	ه	ێ		ی	ۆ	وو	و

(b) vowels

Latin	.	;	,	%	!	?	:		
Arabic	۔	؛	،	٪	!	؟	:	(U+200C)	ـ (U+0640)

(c) punctuation marks

Table 1: A comparison of the two common scripts of Kurdish, Latin-based and Arabic-based

4 Word Boundary in Kurdish

In both the Latin-based and Arabic-based scripts of Kurdish, whitespaces are used for delimiting word boundaries. In addition, the ZWNJ in the Arabic-based script is also commonly used for separating words, particularly verbs that are consisted of more than one word. Having said that, none of these delimiters are deterministic for word boundary in Kurdish (Esmaili, 2012) due to the issues addressed in this section.

4.1 Orthographic Inconsistencies

Despite the efforts within the Kurdish linguistic community to raise awareness regarding orthography and to promote writing guidelines, such as (Hashemi, 2016) for Sorani and (Aydoğan, 2012) for Kurmanji, there is no unified orthography for Kurdish (Ahmadi, 2019). As such, various variations are found with respect to writing a specific word in Kurdish texts. For instance, numbers followed by a morpheme, as in "*di 18ê Adarê*" "on March 18th", may be separated by an apostrophe as in *18'ê*, a hyphen as in *18-ê* or without any punctuation mark.

4.2 Excessive Concatenation

Characters in the Latin script have only one grapheme without changing form. However, depending on the position within the word, characters in the Arabic script may have four different graphemes, namely initial, middle, final and isolated. According to their joining property, characters are categorized into right-joining as ـو , dual-joining as ـبـ and non-joining as digits. This characteristic of the Arabic script may result the reckless concatenation of words without proper spacing. For instance, the word لەوێشدایە (*lewêşdaye*) "(it) is also there" is composed of five words written as one single word, namely لە (*le*) 'from', وێ (*wê*) 'there', ش (*ş*) '*also*', دا (*da*, postposition) and یە (*ye*) 'is'. Such an excessive concatenation creates larger number of tokens represented as one and further complicates the tokenization task.

4.3 Compound Words

Having a relatively few number of around 300 single-word verbs, i.e. verbal lexemes, Kurdish extensively uses compound forms to develop its vocabulary (Walther and Sagot, 2010; Traida, 2007). A compound, also known as multi-word expression (MWE), is a more complex type of word which is consisted of two or more base words. Compound forms represent various challenges in many NLP tasks, including tokenization (Sag et al., 2002; Nasr et al., 2015). Due to the aforementioned issues, finding boundary of compound forms is a non-trivial task as well. In Kurdish, compound forms are written in many different ways, with and without space, using ZWNJ and rarely, using hyphen.

Compound type	Construction	Example		Gloss
		Sorani	Kurmanji	
Verb (V)	particle + V	*řo-nîstin* ڕۆنیشتن	*rû-nîştin*	(over-sit) "to sit down"
	ADJ + V	*germ-kirdin* گەرم- کردن	*germ-kirin*	(warm-do) "to heat"
	N + V	*masî-girtin* ماسی- گرتن	*masî-girtin*	(fish-take) "to fish"
	preposition + V (+ postposition)	*lê-kewtin* لێ- کەوتن	*lê-ketin*	(to that-fall) "to hit"
	V + postposition	*kirdin-ewe* کردنەوە		(do-again) "to open"
	coord. comp. + V	*cê-be-cê-kirdin* جێ-بە-جێ- کردن	*cê-bi-cê-kirin*	(place-to-place-do) "to move"
Noun (N)	infinitive of comp. verbs	*beşdar-bûn* بەشدار-بوون	*beşdar-bûn*	(participate-be) "involvement"
	N + ADJ	*girê-kwêr* گرێ-کوێر	*girê-hişk*	(knot-blind) "hard knot"
	coord. comp.	*cil-û-berg* جل-و-بەرگ	*cil-û-berg*	(cloth-and-cover) "clothes"
	N + present stem	*goranî-bêj* گۆرانی-بێژ	*stran-bêj*	(song-sing.PRS.STEM) "singer"
Adjective (ADJ)	preposition + N	*bê-tam* بێ-تام	*bê-çêj*	(without-taste) "bland"
	coord. comp.	*dûr-û-dirêj* دوور-و-درێژ	*dûr-û-dirêj*	(far-and-long) "detailed"
	ADJ + N	*ciwan-mêr* جوان-مێر	*xweş-mêr*	(young-man) "affable"
Adverb (ADV)	preposition + N	*be-başî* بە-باشی	*bi-başî*	(with-goodness) "nicely"
	preposition + ADJ (+ postposition)	*le-zû-ewe* لە-زوو-ەوە	*ji-zû-ve*	(from-early) "long ago"
	preposition + coord. comp.	*be-lez-û-bez* بە-لەز-و-بەز	*bi-lez-û-bez*	(with-haste-and-race) "hurriedly"
Preposition	preposition + preposition	*be-ser* بە-سەر	*bi-ser*	(with-over) "over"
Conjunction	conjunction + conjunction	*heta-kû* هەتا-کوو	*heta-ku*	(until-that) "so that"

Table 2: Some of the formal constructions of compound (comp.) forms in Kurdish that are consisted of free morphemes. For consistency in writing, composing words are separated by a hyphen

In addition to the verbal compounds, Kurdish widely takes use of coordinative compounds (coord. comp.), i.e. compounds which are formed with conjunction و (*û*) 'and'. Since compound forms carry one meaning, they are usually considered as one token. Moreover, phrasal words are frequently used in Kurdish, such as مردوو-مراو (*mirdû-miraw*) "hapless (adj)"[2] or خوا-خراو-بۆ- کردگ (*xwa-xiraw-bo-kirdig*) "cursed (adj)"[3] respectively in Babani and Ardalani subdialects of Sorani.

Table 2 provides some of the frequent constructions used to create compound forms in Sorani and Kur-

[2]Literally meaning *the one whose dead is dead*
[3]Literally meaning *the one who is cursed by god*

manji Kurdish. It is worth noting that only compound forms which are consisted of free morphemes are provided. Many Kurdish compound forms are produced using inflectional and derivational morphemes which are not covered in this study. In addition, the current formal constructions can be further combined and form more complex compounds, such as دەست-تێ-وەردان (*dest-tê-wer-dan*) 'to manipulate'.

5 Approach

As a preliminary study, we focus on the application of a lexicon of lemmata and morphological analysis for tokenization of Kurdish texts. Moreover, we follow the common practices in tokenization, such as detecting digits, dates, URLs and punctuation marks as distinct tokens. This sub-task is called "normalization prior to tokenization" (Dridan and Oepen, 2012). Given the complexity of detecting word boundaries in Kurdish, particularly in the Arabic-based script of the Sorani dialect, we carry out the task of tokenization based on the syntactic property of the words. In other words, if a sequence of characters, whether delimited by spaces or not, can have a syntactic role, we consider them as a distinct token. Therefore, tokens in Kurdish can be words such as *bira* 'brother', compound words such as *germ kirin* 'to heat (something) up', clitics such as تان *=tan* (2.PL pronominal endoclitic) and affixes such as ەکان *-ekan* (definite plural marker). The following shows an example of how the input sentences in Sorani and Kurmanji are tokenized in our system:

- Sorani
  - Raw: "دوا کەوتنی شێوازەکانی بەرهەمهێنان"
  - Tokenized: ['__دوا- کەوتن__ی_', '_شێواز_ەکان_ی_', '__بەرهەم-هێنان__']
- Kurmanji
  - Raw: "endamên encûmena wezîrên herêma Kurdistanê"
  - Tokenized: `['_endam_ên_', 'encûmen_a', '_wezîr_ên_', '_herêm_a_', '_Kurdistan_ê_']`

5.1 Lexicon

To develop a lexicon for our task, we use the lexicographic material of FreeDicts[4] and the Kurdish Wiktionary, Wîkîferheng[5]. The two resources are available for Sorani and Kurmanji in the Latin-based and Arabic-based scripts of Kurdish. After merging these two resources, we further clean the data by removing the duplicates and normalizing the characters such as diacritical characters '*ř*' and '*ł*'. We also transliterate the Sorani lexicon into the Arabic-based script using Wergor[6] (Ahmadi, 2019). Overall, 8,180 and 9,970 headwords are collected in Sorani and Kurmanji among which 1,513 and 1,507 lemmata are compound forms in Sorani and Kurmanji, respectively.

Using simple regular expressions, the headwords which are consisted of more than one word with a space are retrieved. If the compound form can undergo orthographic inconsistency, the words in such compound forms are separated by a figure dash, i.e. −. Using this special character, we can distinguish the compound forms which can possibly be written differently from the headword in the dictionary. For instance, the words in ولایەتە یەکگرتووەکانی ئەمەریکا (*Wiłayete Yekgirtûwekanî Emerîka*) "the United States of America" are considered to be consistently separated using a whitespace while the space in *birîndar bûn* "to be wounded" can be kept or omitted based on the writer's choice as in *birindar bûn* or *birindarbûn*.

Given that there are various ways to write compound words in Kurdish and due to the lack of a unified orthography, we generate all the possible forms, with and without a whitespace, for each compound entry in our lexicon. The generated forms are then saved in JSON where each compound headword is associated with the possible forms. Listing 1 provides an example of the compound headword *bi-can-û-bên* 'eagerly' in Kurmanji and ئاخر-و-تۆخر 'end' in Sorani and their corresponding forms.

[4] https://freedict.org
[5] https://ku.wiktionary.org
[6] https://github.com/sinaahmadi/wergor

```
{
    "bi-can-û-bên": {
        "token_forms": [
            "bicanûbên",
            "bi canûbên",
            "bican ûbên",
            "bi can ûbên",
            "bicanû bên",
            "bi canû bên",
            "bican û bên",
            "bi can û bên"
]}}
```

Listing 1: A Kurmanji compound lemma and its possible forms in the lexicon in JSON

```
{
  "ئاخر-و-تۆخر": {
        "token_forms": [
                "ئاخرو تۆخر",
                "ئاخروتۆخر",
                "ئاخر و تۆخر",
                "ئاخر وتۆخر",
                "ئاخروو تۆخر",
                "ئاخروتۆخر"
                ]
}
}
```

Listing 2: A Sorani compound lemma and its possible forms in the lexicon in JSON

5.2 Morphological Analyzer

In order to retrieve the lemma form, retrieving inflectional morphemes is required. To do so, we first create a list of the clitics and inflectional affixes used in Kurmanji and Sorani. This list is provided in Table 3. In addition to the given form of the morphemes, some of them can be concatenated and appear in compound forms of two or more morphemes. For instance, the comparative suffix تر (*-tir*) in Sorani can be concatenated with the article maker ەکە (*-eke*) and result the compound form تەرەکە (*-tireke*) as in بەرزتەرەکە *berztireke* 'the higher one'. In addition, some of the clitics appear in an erratic pattern, i.e. depending on their syntactic function they can appear as proclitic, enclitic or endoclitic (Walther, 2012). This is particularly the case of the Sorani pronominal clitics and یش-ش (*îş, ş*). The latter can be translated as 'also, even' and is equivalent to *jî* in Kurmanji (Thackston, 2006). Given such complexities, we create a list of possible forms of the combination of these bound morphemes according to the Kurdish morphology and categorize them based on their position in the word, i.e. before or after the host word. This way, the task of morphological analysis can be carried out in a more simplified way with fewer particular cases to directly consider. The list contains 161 Sorani and 46 Kurmanji compound bound morphemes that can appear after the host and, 11 Sorani and 17 Kurmanji morphemes that can appear before the host. Once a compound morpheme retrieved in a word, it is then replaced by the composing morphemes as in بەرزتەرەکە → بەرز + تەرەکە → بەرز + تر + ەکە (*berz**tireke*** → *berz* + ***tireke*** → *berz* + ***tir*** + ***eke***) where the bound morphemes are shown in bold.

5.3 Tokenization System

Given a text in Kurdish, our tokenization approach is carried out following these steps.

1. **Text preprocessing**: In this step, we unify the encoding of characters, add spaces around punctuation marks, numbers, dates and URLs and remove ZWNJ.
2. **Compound word tokenization**: We append a space before and after the compound lemmas in the lexicon, delimit any occurrence in the text using two ‗ (U+2581). The figure dashes in the compound lemmas are replaced with whitespace. In addition to the lemma, the forms associated with each lemma are to be searched and delimited accordingly.
3. **Word tokenization**: Given the text with compound tokens, we split the text by space and delimit the words that match an entry in the lexicon using one ‗. If a word is not found in the lexicon, we proceed to the next step.
4. **Morphological analysis**: Given the sorted list of the morphemes based on length (longest to shortest), we retrieve the prefixes and suffixes in the word. A prefix or suffix is considered valid only if applying steps 1 and 2 on the remaining of the word returns a match in the lexicon. If so, two tokens consisting of the affix and the word are separated and delimited by a ‗. The affix is accordingly replaced by the composing parts, if any.

This procedure is illustrated in the flowchart in Figure A.4. The output of the tokenization system is a list of tokens. If a word is not delimited throughout this process, it is returned in the original form. Retrieving words based on the their length, which is used in the morphological analysis, is also known as the maximum matching algorithm and has been previously used for the same task (Webster and Kit, 1992). In addition to the word tokenization, we also provide a simple sentence tokenizer using punctuation marks, line breaks, URLs and abbreviations such as هتد (*htd)* in Sorani and *hwd* in Kurmanji for 'etc'.

Description	Morphemes	
	Sorani	Kurmanji
preposition	*le, we, de, ře, be* به - ڕه - ده - وه - له	*ba, berî, beyî, bê, bi, der, di, ji, li, ve*
postposition	*da, řa, ewe, we* وه - ەوه - ڕا - دا	*de, re, ve, da, ra, va*
absolute prepositions and postpositions	*pê, lê, tê, wê, ê* یێ - وێ - تێ - لێ - پێ	*pê - tê - jê - lê*
reciprocal verbal particles	*pêk, řêk, têk, lêk, wêk* لێک - تێک - ڕێک - پێک وێک	*lêk, jêk, pêk, têk*
article marking suffixes	*êk, an, gel, ekan, yekan, ek, yek, eke, yeke, ekan, yekan, ane, e, gele* - ەکان - گەل - ان - یەک - ەکە - یەک - ەک - یەکان - ە - انە - یەکان - ەکان - یەکە گەلە	*ê, î, y, an, ek, yek, ekî, ekê, yekî, yekê, in, ine, inan, ên*
IZAFA	*î, y, e* / ی - ە	*ê, a, yê, ya, yên*
locative and vocative suffixes	*îne, o, ê, yê* ینە - یێ - ۆ - ینە	*o, ê, ên, no*
pronominal cliticcs	*im,m,it,t,man,tan,yan* یان - تان - مان - ی - ت - م	*min, te, wî, wê, vî, vê, me, we, wan, van*
present copula	*im,m,î,y,ît,e,ye,în,in,n* ن - ین - یە - ە - یت - ی - م	*im, î, e, in, me, yî, ye, ne*
superlative and comparative suffixes	*tir, tirîn* / ترین - تر	*tir, tirîn*
other endoclitics	*îş, ş* / یش - ش	

Table 3: The morphemes used in our morphological analyzer to extract tokens from word forms. The morphemes in the Latin-based and Arabic-based alphabets are respectively separated by a comma and a hyphen

6 Experiments

6.1 Data Annotation

In order to evaluate the performance of our tokenization system, we create a gold-standard dataset by annotating 100 sentences from the KTC corpus (Abdulrahman et al., 2019) for Sorani and 100 sentences from the PEWAN corpus (Esmaili et al., 2013) for Kurmanji. In the annotation process, we followed the same guidelines regarding the definition of tokens in Kurdish depicted in this study. The datasets are available in the Text Corpus Format (TCF) (Heid et al., 2010) and can be further enriched by adding annotations regarding lemmata, part-of-speech and constituent parse trees. Table 4 provides basic statistics of the annotated datasets and the samples of two Sorani and Kurmanji sentences are provided in Figure A.2.

Dataset	# sentences	# word types	# space-delimited words	# annotated tokens
Kurmanji	100	727	1378	2066
Sorani	100	904	1201	1994

Table 4: Statistics of the annotated datasets for the evaluation of the tokenization system. # denotes the number

6.2 Tokenization Models

We create our baseline model using the `WordPunct` tokenizer of NLTK (Loper and Bird, 2002). This technique tokenizes text into a sequence of alphabetic and non-alphabetic characters using the regular expression "`\w+|[^\w\s]+`". In addition, we train a few unsupervised neural models using HuggingFace Tokenizers[7] and SentencePiece[8] (Wu et al., 2016). In the first case, we use `WordPiece` which is a subword tokenization algorithm used for BERT language model (Devlin et al., 2018). In the latter, we train tokenization models using Byte Pair Encoding (`BPE`) (Sennrich et al., 2016), unigram language model (`Unigram`) (Kudo, 2018) and `Word` model types. It is worth noting that the `Word` tokenization model is in fact a language model trained on data pre-tokenized with the `WordPunct` tokenizer.

The unsupervised neural models are trained with various vocabulary sizes and a character coverage of 1.0 using the available Sorani Kurdish raw corpora, namely Pewan corpus containing 18M Sorani words and 4M Kurmanji words (Esmaili et al., 2013), the Kurdish Textbooks Corpus (KTC) containing 693K Sorani words (Abdulrahman et al., 2019), Veisi et al. (2020)'s Sorani corpus containing 8.1M words and the Sorani Kurdish folkloric lyrics corpus containing 49K words (Ahmadi et al., 2020). Due to the limited size of the Kurmanji data, we also used the raw text of the Kurmanji Wikipedia containing 3M words[9]. The corpora are preprocessed by unifying character encoding according to the alphabets in Table 1.

6.3 Evaluation Metrics

The performance of tokenization methods is more meaningful to be evaluated within an end-to-end scenario such as machine translation and syntactic parsing (Resnik and Lin, 2010, p 275). Due to the limited advances in Kurdish language processing, we evaluate our tokenization as a component alone. To this end, we calculate accuracy (acc) by comparing the gold-standard tokens versus the output of each system and dividing the number of correctly-tokenized tokens by the whole number of tokens. In addition to this overall accuracy, we evaluate the accuracy of the systems with respect to compound words ($acc_{comp.}$) where only the compound lemmata in the lexicon is evaluated.

In addition to accuracy, we use the Bilingual Evaluation Understudy Score, more commonly known as BLEU (Papineni et al., 2002). This scoring method is widely used for evaluation purposes in machine translation and has also many applications in evaluating the quality of a generated text with comparison to a reference one. In our case, we calculate the cumulative n-gram case-sensitive BLEU score (BLEU-n) (Yang et al., 2018) on the gold-standard tokens and the output of the various tokenization methods. The cumulative n-gram considers individual n-gram scores from 1 to n, in our case 4 and using their weighted geometric mean, calculates the overall BLEU score. This way, the performance of single-word tokens, i.e. BLEU-1, as well as compound words are taken into account.

6.4 Results

Table 5 presents the evaluation results of the unsupervised models with comparison to the baseline and our approach. In all the models, the BLEU scores decrease gradually from BLEU-1 to BLEU-4. This indicates that the models perform relatively better with respect to the tokenization of words which are composed of one single token and are accompanied with few morphemes while, a compound word with richer morphological form is more challenging to be tokenized correctly. On the other hand, by increasing the vocabulary size, the overall accuracy, i.e. acc increases in most cases. Figure 1 illustrates this

[7] `https://github.com/huggingface/tokenizers`
[8] `https://github.com/google/sentencepiece`
[9] Based on May 2020 dump

Model type	#Vocab.	Sorani						Kurmanji					
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	*acc* %	$acc_{comp.}$ %	BLEU-1	BLEU-2	BLEU-3	BLEU-4	*acc* %	$acc_{comp.}$ %
BPE	4000	0.94	0.85	0.77	0.7	7.78	20.04	0.93	0.85	0.77	0.71	7.07	12.03
	8000	0.97	0.89	0.83	0.77	12.89	35.63	**0.98**	0.92	0.87	0.82	15.28	25.97
	16000	**0.98**	0.92	0.86	0.81	13.68	50.03	0.97	0.92	0.88	0.84	17.19	39.59
	32000	0.97	0.92	0.87	0.83	14.87	62.77	0.96	0.92	0.89	0.85	18.82	50.83
Unigram	4000	0.95	0.88	0.82	0.76	11.55	24.29	0.94	0.86	0.8	0.74	10.89	13.62
	8000	0.97	0.92	0.87	0.82	14.38	40.15	**0.98**	0.93	0.89	0.84	15.71	30.4
	16000	**0.98**	0.93	0.89	0.85	15.37	51.36	0.97	0.93	0.89	0.86	18.62	42.89
	32000	0.97	0.93	0.9	0.86	17.3	62.64	0.96	0.93	0.9	**0.87**	19.39	51.88
Word	4000	0.88	0.83	0.79	0.74	5.75	81.22	0.91	0.87	0.84	0.81	8.21	87.64
	8000	0.89	0.84	0.8	0.76	6.1	92.5	0.92	0.88	0.85	0.82	8.55	89.95
	16000	0.89	0.85	0.81	0.77	6.59	94.89	0.92	0.88	0.85	0.82	9.12	92.2
	32000	0.9	0.85	0.81	0.77	6.35	**96.35**	0.92	0.89	0.86	0.82	9.12	**94.32**
WordPiece	4000	0.94	0.86	0.78	0.71	9.17	17.32	0.85	0.80	0.75	0.70	13.94	19.43
	8000	0.93	0.87	0.8	0.75	13.29	34.04	0.83	0.79	0.75	0.71	13.90	34.96
	16000	0.92	0.87	0.82	0.77	13.73	47.78	0.81	0.78	0.75	0.72	13.09	47.65
	32000	0.9	0.86	0.82	0.78	13.29	59.99	0.81	0.78	0.75	0.72	12.13	58.63
WordPunct (baseline)		0.93	0.91	0.88	0.85	9.72		0.95	0.92	0.9	**0.87**	15.09	
Our system		**0.98**	**0.95**	**0.91**	**0.87**	**30.44**		0.97	**0.94**	**0.91**	**0.87**	**31.38**	

Table 5: Evaluation of various unsupervised methods and our tokenization system

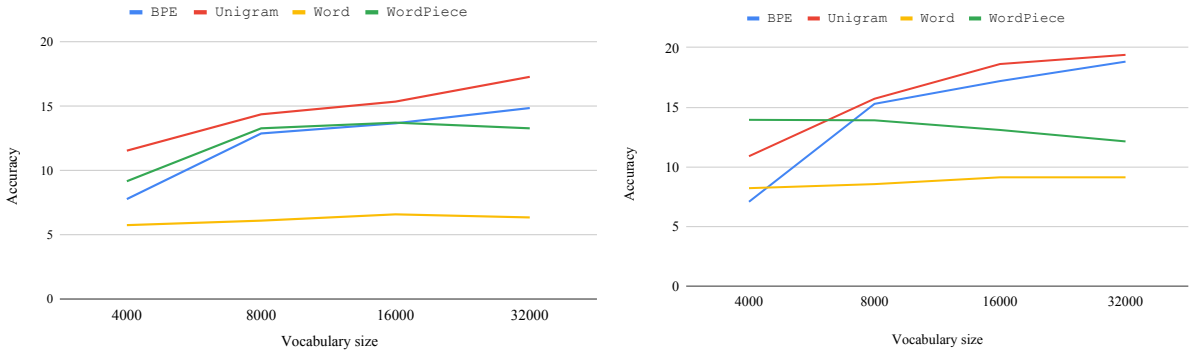

Figure 1: Accuracy of the unsupervised tokenization models in Sorani (left) and Kurmanji (right)

correlation in the various models. In almost all cases, the results of our system outperforms the other methods with a remarkable difference in the accuracy. It is worth mentioning that the accuracy of the baseline with respect to compound forms, i.e. $acc_{comp.}$ is either 100% or 0% depending on adding a whitespace between composing parts or not. Figure A.3 presents an example of the output of the models.

7 Conclusion and Future Work

In this paper, we presented a tokenization system for the Sorani and Kurmanji dialects of Kurdish. Having a complex morphology and various compound form constructions, Kurdish represents non-trivial challenges to the tokenization task. Using a lexicon and a morphological analyzer, our system outperforms unsupervised neural methods and can also be used to detect compound forms efficiently.

One limitation of the current study is the tokenization of compound verbs. In addition to tense, aspect, person and mood, verbs are inflected according to the patient, i.e. object of transitive verbs, and can be accompanied by other affixes such as ەوە (*-ewe*) to indicate repetition and ش/یش (*=îş/=ş*) to indicate emphasis. Some clitics appear within the root of the verb, therefore called endoclitic, and create more complex forms. If in the tokenization task, such parts of the verbs are split into tokens, the compound verb is also split into its composing parts instead of being tokenized as one.

Given the relatedness of lemmatization to the current task, we believe that extending the current study can be beneficial to create a lemmatization system for Kurdish as well. Moreover, enriching the lexicons, particularly by including further compound form constructions, should also be considered in the future by integrating further collaboratively-curated open resources.

Acknowledgements

The author would like to thank the four anonymous reviewers for their constructive comments.

References

Roshna Omer Abdulrahman, Hossein Hassani, and Sina Ahmadi. 2019. Developing a Fine-Grained Corpus for a Less-resourced Language: the case of Kurdish. *arXiv preprint arXiv:1909.11467*.

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

Sina Ahmadi, Hossein Hassani, and Kamaladdin Abedi. 2020. A Corpus of the Sorani Kurdish Folkloric Lyrics. In *Proceedings of the 1st Joint Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Workshop at the 12th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

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

Mustafa Aydoğan. 2012. *Rêbera rastnivîsînê*. Weşanxaneya Rûpelê. Ziman. Rûpel.

Shuanhu Bai, Horng Jyh Paul Wu, Haizhou Li, and Gareth Loudon. 2001. System for Chinese tokenization and named entity recognition, October 30. US Patent 6,311,152.

Michael L Chyet and Martin Schwartz. 2003. *Kurdish-English Dictionary*. Yale University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Miguel Domingo, Mercedes Garcıa-Martınez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2018. How Much Does Tokenization Affect Neural Machine Translation? *arXiv preprint arXiv:1812.08621*.

Rebecca Dridan and Stephan Oepen. 2012. Tokenization: Returning to a long solved problem—a survey, contrastive experiment, recommendations, and toolkit—. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382.

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for Sorani Kurdish. In *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*, pages 1–7. IEEE.

Kyumars Sheykh Esmaili. 2012. Challenges in Kurdish text processing. *arXiv preprint arXiv:1212.0074*.

Ismaïl Kamandâr Fattah. 2000. *Les dialectes kurdes méridionaux: étude linguistique et dialectologique*. Acta Iranica : Encyclopédie permanente des études iraniennes. Peeters.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Martin Forst and Ronald M Kaplan. 2006. The importance of precise tokenizing for deep grammars. In *LREC*, pages 369–372.

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

Jin Guo. 1997. Critical tokenization and its properties. *Computational Linguistics*, 23(4):569–596.

Benoit Habert, Gilles Adda, Martine Adda-Decker, P Boula de Marëuil, Serge Ferrari, Olivier Ferret, Gabriel Illouz, and Patrick Paroubek. 1998. Towards tokenization evaluation. In *Proceedings of LREC*, volume 98, pages 427–431.

Dyako Hashemi. 2016. Kurdish orthography [In Kurdish]. `http://yageyziman.com/Renusi_Kurdi.htm`. Accessed: 2020-07-25.

Martin Haspelmath and Andrea D Sims. 2013. *Understanding morphology*. Routledge.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A Corpus Representation Format for Linguistic Web Services: The D-SPIN Text Corpus Format and its Relationship with ISO Standards. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Tatsuya Hiraoka, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Stochastic tokenization with a language model for neural text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1620–1629.

Ronald M Kaplan. 2005. A method for tokenizing text. *Inquiries into words, constraints and contexts*, 55.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational linguistics*, 32(4):485–525.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Ludovic Lebart, André Salem, and Lisette Berry. 1997. *Exploring textual data*, volume 4. Springer Science & Business Media.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Yaron Matras. 2017. Revisiting Kurdish dialect geography: Preliminary findings from the Manchester Database. `http://kurdish.humanities.manchester.ac.uk/wp-content/uploads/2017/07/PDF-Revisiting-Kurdish-dialect-geography.pdf`. [Online; accessed 02-Aug-2020].

Yaron Matras. 2019. Revisiting Kurdish dialect geography: Findings from the Manchester Database. *Current issues in Kurdish linguistics*, 1:225.

Tony McEnery and Andrew Wilson. 2003. Corpus linguistics. *The Oxford handbook of computational linguistics*, pages 448–463.

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information retrieval*, 7(1-2):73–97.

Alexis Nasr, Carlos Ramisch, José Deulofeu, and Andre Valli. 2015. Joint Dependency Parsing and Multiword Expression Tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1116–1126.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tarik A Rashid, Arazo M Mustafa, and A Saeed. 2017. A robust categorization system for Kurdish Sorani text documents. *Inf. Technol. J*, 16(1):27–34.

Zobia Rehman, Waqas Anwar, Usama Ijaz Bajwa, Wang Xuan, and Zhou Chaoying. 2013. Morpheme matching based text tokenization for a scarce resourced language. *PloS one*, 8(8):e68178.

Philip Resnik and Jimmy Lin. 2010. Evaluation of NLP Systems. *The handbook of computational linguistics and natural language processing*, 57.

Siamak Rezaie. 2001. Tokenizing an Arabic script language. *Arabic language processing: Status and prospects, ACL/EACL*.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection. In *KONVENS*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Mehrnoush Shamsfard, Soheila Kiani, and Yaseer Shahedi. 2009. STeP-1: standard text preparation for Persian language. In *Third Workshop on Computational Approaches to Arabic Script-based Languages*, pages 859–865.

Givi Tavadze. 2019. Spreading of the Kurdish Language Dialects and Writing Systems Used in the Middle East. *Bull. Georg. Natl. Acad. Sci*, 13(1).

Wheeler M. Thackston. 2006. *Kurmanji Kurdish:-A Reference Grammar with Selected Readings*. Harvard University.

Sandrine Traida. 2007. *Morphosyntactic Study of the compound verbs in Sorani Kurdish Étude morpho-syntaxique des verbes composés (nom-verbe) en kurde (dialecte sorani) [in French]*. PhD thesis at the Université Paris 3 - Sorbonne Nouvelle.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2020. Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

Benjamin Waldron, Ann A Copestake, Ulrich Schäfer, and Bernd Kiefer. 2006. Preprocessing and Tokenisation Standards in DELPH-IN Tools. In *LREC*, pages 2263–2268.

Géraldine Walther and Benoît Sagot. 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)*.

Géraldine Walther. 2012. Fitting into morphological structure: accounting for Sorani Kurdish endoclitics. In *Mediterranean Morphology Meetings*, volume 8, pages 299–321. [Online; accessed 02-Aug-2020].

Jonathan J Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

An Yang, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li. 2018. Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 98–104.

Nasser Zalmout and Nizar Habash. 2017. Optimizing tokenization choice for machine translation across multiple target languages. *The Prague Bulletin of Mathematical Linguistics*, 108(1):257–269.

A Appendix

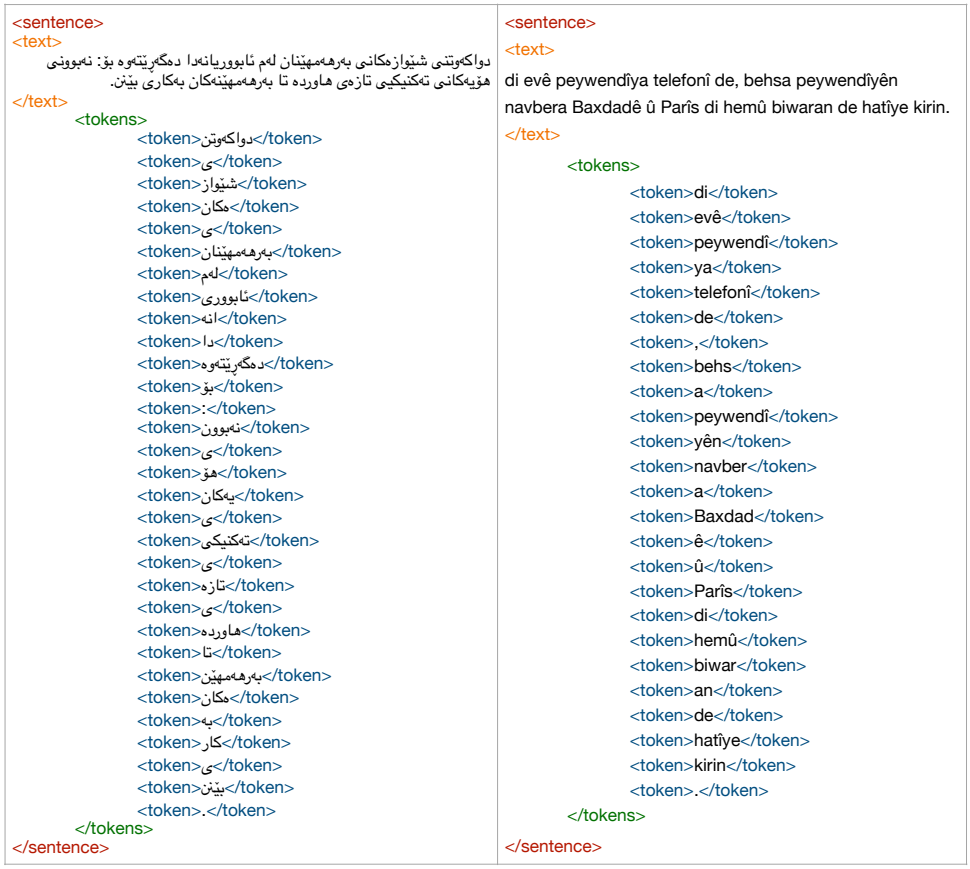

Figure A.2: An example of the annotated tokens of two different sentences in Sorani (left) and Kurmanji (right) in the Text Corpus Format

Reference	دواکەوتن ی شێواز ەکان ی بەرهەمهێنان لەم ئابووری انە دا دەگەڕێتەوە بۆ : نەبوون ی هۆ یەکان ی تەکنیکی ی تازە ی هاوردە تا بەرهەمهێن ەکان بە کار ی بێنن .
Our system	دواکەوتن ی شێواز ەکان ی بەرهەمهێنان لەم ئابووریانەدا دەگەڕێتەوە بۆ : نەبوونی هۆیەکان ی تەکنیکی تازە ی هاوردە تا بەرهەمهێنەکان بەکاری بێن ن .
BPE	دواکەوتنی شێوازەکانی بەرهەمهێنان لەم ئابووری انەدا دەگەڕێتەوە بۆ : نەبوونی هۆیەکانی تەکنیک یی تازەی هاوردە تا بەرهەمهێن ەکان بەکاری بێنن .
Unigram	دواکەوتنی شێوازەکانی بەرهەمهێنان لەم ئابووری انەدا دەگەڕێتەوە بۆ : نەبوونی هۆیەکانی تەکنیکی ی تازەی هاوردە تا بەرهەمهێن ەکان بەکاری بێنن .
Word	دواکەوتنی شێوازەکانی بەرهەمهێنان لەم ئابووریانەدا دەگەڕێتەوە بۆ: نەبوونی هۆیەکانی تەکنیکیی تازەی هاوردە تا بەرهەمهێنەکان بەکاری بێنن.
WordPiece	دواکەوتنی شێوازەکانی بەرهەمهێنان لەم دەگەڕێتەوە بۆ : نەبوونی هۆیەکانی تەکنیکی ی تازەی هاوردە تا بەرهەمهێن ەکان بەکاری بێنن .
WordPunct	دواکەوتنی شێوازەکانی بەرهەمهێنان لەم ئابووریانەدا دەگەڕێتەوە بۆ : نەبوونی هۆیەکانی تەکنیکیی تازەی هاوردە تا بەرهەمهێنەکان بەکاری بێنن .

(a) Sorani

Reference	di evê peywendî ya telefonî de , behsa peywendî yên navber a Baxdad ê û Parîs di hemû biwaran de hatîye kirin .
Our system	di ev ê peywend îya telefonî de , behs a peywendî yên navbera Baxdadê û Parîs di hemû biwaran de hatîye kirin .
BPE	di evê peywendîya telefonî de , behsa peywendîyên navbera Baxdadê û Parîs di hemû biwaran de hatîye kirin .
Unigram	di evê peywendîya telefonî de , behsa peywendîyên navbera Baxdadê û Parîs di hemû biwaran de hatîye kirin .
Word	di evê peywendîya telefonî de, behsa peywendîyên navbera Baxdadê û Parîs di hemû biwaran de hatîye kirin.
WordPiece	di evê peywendîya telefonî de , behsa peywendîyên navbera û di hemû biwaran de hatîye kirin .
WordPunct	di evê peywendîya telefonî de , behsa peywendîyên navbera Baxdadê û Parîs di hemû biwaran de hatîye kirin .

(b) Kurmanji

Figure A.3: The output of the unsupervised neural tokenization models with vocabulary size 32000, the baseline (`WordPunct`) and our system

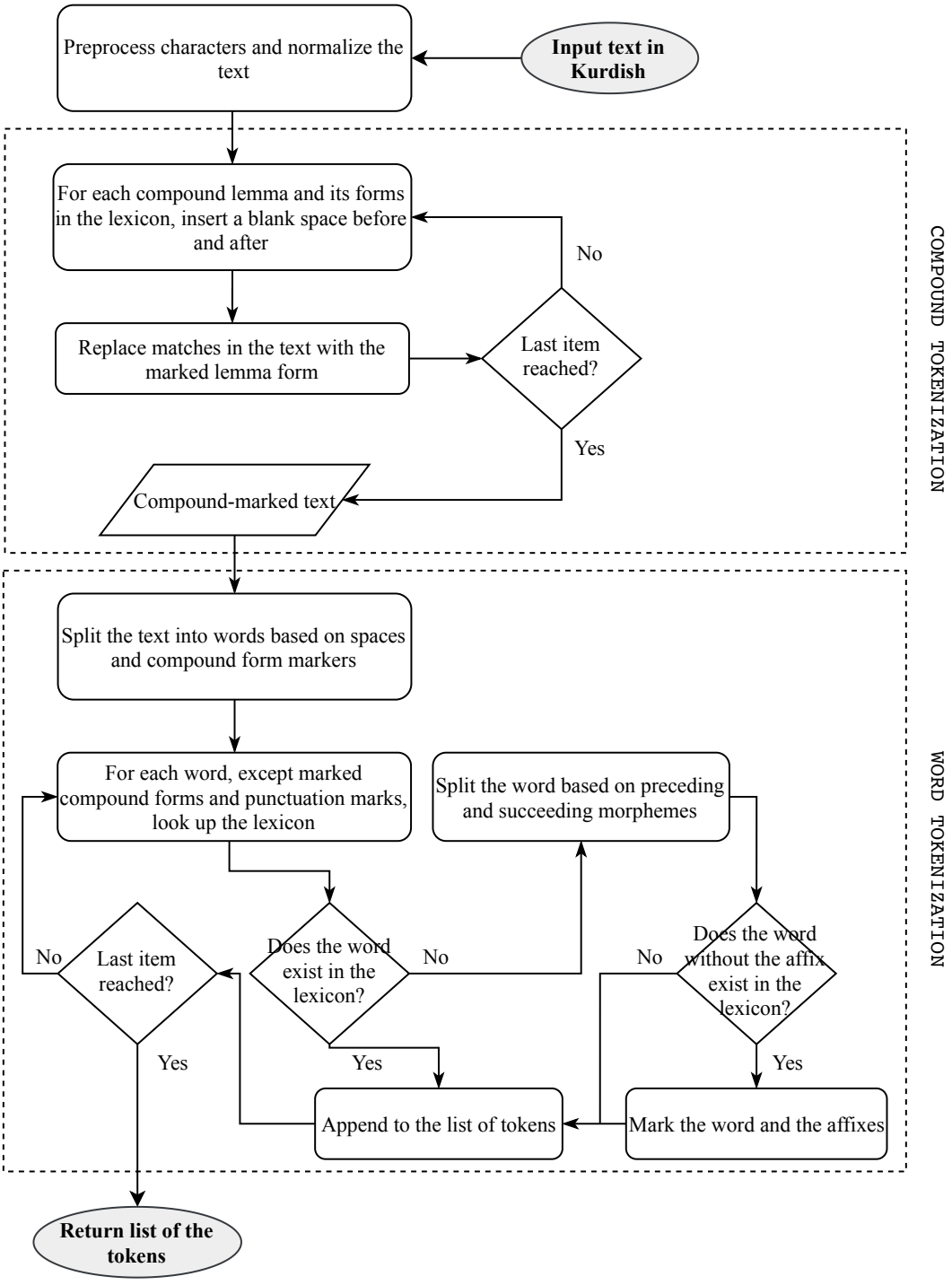

Figure A.4: The flowchart of the Kurdish tokenization system proposed in this paper. Marking action refers to appending __ (U+2581) before and after a token