Creating a Fine-Grained Corpus for a Less-resourced Language: the case of Kurdish

Roshna Omer Abdulrahman
University of Kurdistan Hewlêr
Kurdistan Region - Iraq
roshna.abdulrahman@ukh.edu.krd

Hossein Hassani
University of Kurdistan Hewlêr
Kurdistan Region - Iraq
hosseinh@ukh.edu.krd

Sina Ahmadi
Insight Centre for Data Analytics
National University of Ireland Galway
Galway - Ireland
sina.ahmadi@insight-centre.org

Abstract

Kurdish is a less-resourced language consisting of different dialects written in various scripts. Approximately 30 million people in different countries speak the language. The lack of corpora is one of the main obstacles in Kurdish language processing. In this paper, we present KTC–the Kurdish Textbooks Corpus, which is composed of 31 K-12 textbooks in Sorani dialect. The corpus is normalized and categorized into 12 educational subjects containing 693,800 tokens (110,297 types). Our resource is publicly available for non-commercial use under the CC BY-NC-SA 4.0 license[1].

1 Introduction

Kurdish is an Indo-European language mainly spoken in central and eastern Turkey, northern Iraq and Syria, and western Iran. It is a less-resourced language (Salavati and Ahmadi, 2018), in other words, a language for which general-purpose grammars and raw internet-based corpora are the main existing resources. The language is spoken in five main dialects, namely, Kurmanji (aka Northern Kurdish), Sorani (aka Central Kurdish), Southern Kurdish, Zazaki and Gorani (Haig and Öpengin, 2014).

Creating lexical databases and text corpora are essential tasks in natural language processing (NLP) development. Text corpora are knowledge repositories which provide semantic descriptions of words. The Kurdish language lacks diverse corpora in both raw and annotated forms (Esmaili et al., 2013; Hassani, 2018). According to the literature, there is no domain-specific corpus for Kurdish.

In this paper, we present KTC, a domain-specific corpus containing K-12 textbooks in Sorani. We consider a domain as a set of related concepts, and a domain-specific corpus as a collection of documents relevant to those concepts (Mason, 2004). Accordingly, we introduce KTC as a domain-specific corpus because it is based on the textbooks which have been written and compiled by a group of experts, appointed by the Ministry of Education (MoE) of the Kurdistan Region of Iraq, for educational purposes at the K-12 level. The textbooks are selected, written, compiled, and edited by experts in each subject and also by language editors based on a unified grammar and orthography. This corpus was initially collected as an accurate source for developing a Sorani Kurdish spellchecker for scientific writing. KTC contains a range of subjects, and its content is categorized according to those subjects. Given the accuracy of the text from scientific, grammatical, and orthographic points of view, we believe that it is also a fine-grained resource. The corpus will contribute to various NLP tasks in Kurdish, particularly in language modeling and grammatical error correction.

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

In the rest of this paper, Section 2 reviews the related work, Section 3 presents the corpus, Section 4 addresses the challenges in the project and, Section 5 concludes the paper.

2 Related work

Although the initiative to create a corpus for Kurdish dates back to 1998 (Gautier, 1998), efforts in creating machine-readable corpora for Kurdish are recent. The first machine-readable corpus for Kurdish is the Leipzig Corpora Collection which is constructed using different sources on the Web (Biemann et al., 2007). Later, Pewan (Esmaili et al., 2013) and Bianet (Ataman, 2018) were developed as general-purpose corpora based on news articles. Kurdish corpora are also constructed for specific tasks such as dialectology (Malmasi, 2016; Hassani, 2018), machine transliteration (Ahmadi, 2019), and part-of-speech (POS) annotation (Walther and Sagot, 2010; Walther et al., 2010). However, to the best of our knowledge, currently, there is no domain-specific corpus for Kurdish dialects.

3 The Corpus

KTC is composed of 31 educational textbooks published from 2011 to 2018 in various topics by the MoE. We received the material from the MoE partly in different versions of Microsoft Word and partly in Adobe InDesign formats. In the first step, we categorized each textbook based on the topics and chapters. As the original texts were not in Unicode, we converted the content to Unicode. This step was followed by a pre-processing stage where the texts were normalized by replacing zero-width-non-joiner (ZWNJ) (Esmaili et al., 2013) and manually verifying the orthography based on the reference orthography of the Kurdistan Region of Iraq. In the normalization process, we did not remove punctuation and special characters so that the corpus can be easily adapted our current task and also to future tasks where the integrity of the text may be required.

Module title	Course level	#Chapters	#Tokens	#Sentences
Economics	12	7	32,823	1,023
Genocide	10	8	16,243	670
Geography	10	10	27,999	884
History	10,12	20	79,845	2,065
Human Rights	10	5	11,527	340
Kurdish	7,8,9,10,12	86	153,334	6,348
Kurdology	10,11 (i)	6	34,282	931
Philosophy	11	6	21,953	549
Physics	1,2,3,4 (i)	30	111,032	4,022
Theology	1,4,5,6,7,8,9,10,11,12	191	115,349	3,661
Sociology	8,9	42	68,044	2,082
Social Study	10	6	21,369	578
Total	31	417	693,800	23,153

Table 1: Statistics of the corpus - In the Course Level column, (i) represents Institute[2].

As an experiment, we present the top 15 most used tokens of the textbooks in KTC, which are illustrated in Figure 1. We observe that the most frequent tokens such as ((economics)ئابووری, (business)بازرگانی) in economics, (= ,× (energy)وزەی) in physics, and ((god)خودای ,(great)گەورە ,(meaning)واتە) in theology are conjunctions, prepositions, pronouns or punctuation. These are not descriptive of any one subject, while each subject's top tokens are descriptive of its content. The plot in Figure 1 follows Zipf's law to some extent, wherein the frequency of a word is proportional to its rank (Powers, 1998). Here, not only the words but also the punctuation and special characters are also considered tokens (see Section 1).

[2] The students could choose to go to the Institutes instead of High Schools after the Secondary School. The Institutes focus on professional and technical education aiming at training technicians.

The corpus is available at `https://github.com/KurdishBLARK/KTC`.[3]

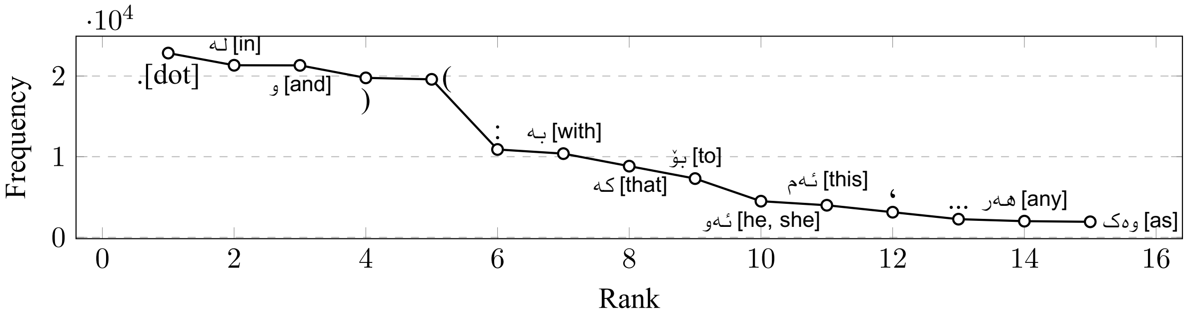

Figure 1: Common tokens among textbook subjects.

4 Challenges

Previously, researchers have addressed the challenges in Kurdish corpora development (Esmaili et al., 2013; Aliabadi et al., 2014; Hassani, 2018). We highlight two main challenges we faced during the KTC development. First, most of the written Kurdish resources have not been digitized (Hassani et al., 2019), or they are either not publicly available or are not fully convertible. Second, Kurdish text processing suffers from different orthographic issues (Ahmadi, 2019) mainly due to the lack of standard orthography and the usage of non-Unicode keyboards. Therefore, we carried out a semi-automatic conversion, which made the process costly in terms of time and human assistance.

5 Conclusion

We presented KTC–the Kurdish Textbook Corpus, as the first domain-specific corpus for Sorani Kurdish. This corpus will pave the way for further developments in Kurdish language processing. We have mad the corpus available at `https://github.com/KurdishBLARK/KTC` for non-commercial use. We are currently working on a project on the Sorani spelling error detection and correction. As future work, we are aiming to develop a similar corpus for all Kurdish dialects, particularly Kurmanji.

Acknowledgements

We would like to appreciate the generous assistance of the Ministry of Education of the Kurdistan Region of Iraq, particularly the General Directorate of Curriculum and Printing, for providing us with the data for the KTC corpus. Our special gratitude goes to Ms. Namam Jalal Rasheed and Mr. Kawa Omer Muhammad for their assistance in making the required data available and resolving of the copyright issues.

References

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

Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards building kurdnet, the kurdish wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 1–6.

Duygu Ataman. 2018. Bianet: A parallel news corpus in turkish, kurdish and english. *arXiv preprint arXiv:1805.05095*.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, 2007.

[3] The materials of KTC are copyrighted. For more information refer to the provided link.

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 *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Gérard Gautier. 1998. Building a kurdish language corpus: An overview of the technical problems. *Proceedings of ICEMCO*.

Geoffrey Haig and Ergin Öpengin. 2014. Introduction to special issue-kurdish: A critical research overview. *Kurdish Studies*, 2(2):99–122.

Hossein Hassani, Emir Turajlić, and Kemal Taljanović. 2019. Digital humanities readiness assessment framework: Dhuraf. *arXiv preprint arXiv:1902.06532*.

Hossein Hassani. 2018. Blark for multi-dialect languages: towards the kurdish blark. *Language Resources and Evaluation*, 52:625–644.

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

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.

David MW Powers. 1998. Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pages 151–160. Association for Computational Linguistics.

Shahin Salavati and Sina Ahmadi. 2018. Building a lemmatizer and a spell-checker for sorani kurdish. *arXiv preprint arXiv:1809.10763*.

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, Benoît Sagot, and Karën Fort. 2010. Fast development of basic nlp tools: Towards a lexicon and a pos tagger for Kurmanji Kurdish. In *International conference on lexis and grammar*, page 0.