Creating a Fine-Grained Corpus for a Less-resourced Language

The Case of Kurdish

Roshna Omer[1],Hossein Hassani[1], Sina Ahmadi[2]
[1]University of Kurdistan Hewlêr, KRG, Iraq [2]Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

Abstract

We present KTC–the Kurdish Textbooks Corpus, which is composed of 31 K-12 textbooks published from 2011 to 2018 in the Sorani dialect of the Kurdish language.The corpus is normalized and categorized into 12 educational subjects containing 693,800 tokens (110,297 types). We did not remove punctuation and special characters so that the corpus can be easily adapted to future tasks where the integrity of the text may be required.

The KTC is publicly available at `https://github.com/KurdishBLARK/KTC`.

The Kurdish Language

Kurdish is an Indo-European language (Salavati and Ahmadi, 2018) That is considered as less-resourced, it consists of different dialects written in various scripts. Approximately 30 million people speak the Kurdish language in different countries mainly in:

- Central and eastern Turkey
- Northern Iraq
- Northern Syria
- Western Iran

Kurdish is spoken in five main dialects (Hassani, 2018):

- Kurmanji (aka Northern Kurdish)
- Sorani (aka Central Kurdish)
- Southern Kurdish
- Zazaki
- Gorani

KTC

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 KTC. The corpus consists of 31 Kurdish Sorani textbooks that were published from 2011 to 2018. KTC is categorized into 12 educational subjects containing 693,800 tokens (110,297 types). In the Course Level column, (i) represents Institute.

Figure

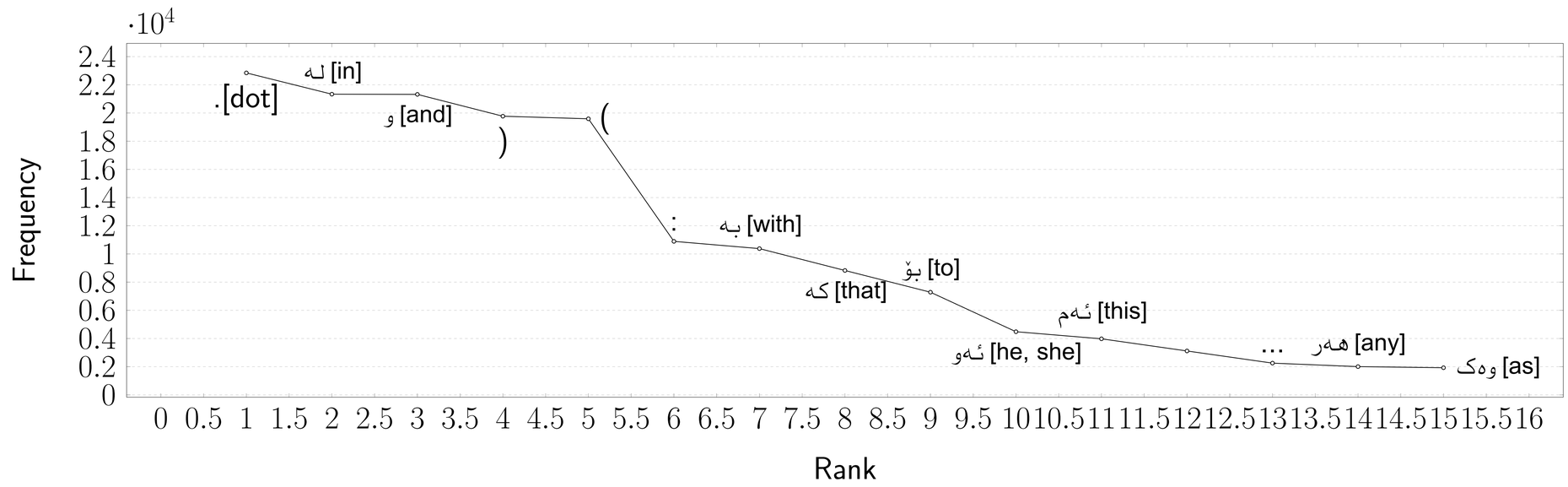

Figure 1: Common tokens among textbook subjects.

Discussion

KTC is categorized based on topics and chapters (see Table 1). We normalized the content by converting it to Unicode and replacing zero-width-non-joiner (ZWNJ) Esmaili et al. (2013).

We present the top 15 most used tokens of the textbooks in KTC, which are illustrated in Figure 1.

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.

Other Kurdish corpora such as 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). However, to the best of our knowledge, currently, there is no domain-specific corpus for Kurdish dialects.

References

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

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

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

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

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

Powers, D. M. (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.

Salavati, S. and Ahmadi, S. (2018). Building a lemmatizer and a spell-checker for sorani kurdish. *arXiv preprint arXiv:1809.10763*.

www.ukh.edu.krd,www.insight-centre.org {hosseinh,roshna.abdulrahman}@ukh.edu.krd,sina.ahmadi@insight-centre.org