https://proceedings.undip.ac.id/index.php/icocas/index ISSN: 3031-5794

A Corpus-Based Analysis of 'Muslim' in the LCC Indonesia 2023

Prihantoro, Adinda Maghfiroh, Windy Arum Sari, Siti Rahmawati Hidayah, Rizki Dwi Nuradita, Ronald Dharmawan

Department of Linguistics, Faculty of Humanities, Diponegoro University, Semarang 50275, Indonesia
prihantoro@live.undip.ac.id

Abstract. Some scholars have visited the representation of Muslims in various English corpora. Unlike previous studies, we here focus on how 'muslim' is represented in LCC 2023, an Indonesian Corpus in CQPweb Lancaster. We aim to (1) identify strong collocates of 'muslim', (2) profile the diachronic distribution of collocates over years (3) semantically categorize them based on Baker et al.'s (2013) classification. Collocates were obtained using Dice's Coefficient and ranked from high to low. As many as 50 of the strongest noun collocates per year. We exported the collocates to identify recurring collocate patterns. Semantic annotations were performed to categorise each collocate. We discover both positive and negative sentiment collocates used to represent 'muslim' in the corpus. This shows how 'muslim' is portrayed in the corpus with both positive and negative images.

Keywords: Collocation, Collocates, Muslim, Corpus Linguistics, Semantic Categorization

1 Introduction

The representation of Muslim groups in public discourse has been studied in different projects, such as Jaszczyk-Grzyb et al. [1], Lestari and Prihantoro [2], and Al-Qattan & Abumiera [3], Knoblock [4], among others. Unlike these studies, which approach the representation of 'muslim' from media perspectives, Baker et al. [5] and Rhinehart [6] adopt corpus approaches to study the representation of muslim. However, they share the same target language as the studies mentioned earlier, English. Their findings highlight negative sentiments towards muslim. Arifa et al. [7], unlike those studies, targets the Indonesian. She discovered that the sentiments are positive. Note that she studied a small corpus built from Jawa Pos media. In this study, we also target Indonesian, but in a much larger and representative corpus, LCC Indonesian 2023 in CQPweb Lancaster. Our Research Questions (RQs) are: 1) What are the strong collocates of 'muslim'? 2) How do collocates progress diachronically?, 3) What are the semantic categorisation of the collocates?

https://proceedings.undip.ac.id/index.php/icocas/index ISSN: 3031-5794

2 Methods

We used corpus approaches, and in particular collocation, to analyse the corpus. The same approach has been implemented by Lestari and Prihantoro [2], Prihantoro ang Gillings [8], Prihantoro ([9],[10]) among others. The target corpus in this study is LCC Indonesian 2023 in CQPweb Lancaster[1], in an amount of more than 500 million word tokens. We used the orthographic search 'muslim' to query the corpus. This is because, unlike English, there is no inflectional variant of 'muslim' in Indonesian. Therefore, it is unnecessary to use a lemma search. To answer RQ1, we used CQPweb's collocation function. We preserved all the default collocation parameters, except the collocation statistic. The Dice coefficient collocation measure was chosen due to its nature as a compromise between the size effect and significance measures. The collocates generated by the system have already been ranked by collocation statistic score from high to low. For RQ2, the same process was repeated, but we applied it on a restricted search to obtain collocates on a yearly basis. We then organise the collocates in a spreadsheet to check whether some collocates were more consistent. The consistency is presented in proportion (%). The higher the proportion, the more consistent a collocate appears across years. For RQ3, we categorised collocates based on Baker et al.'s [5] classification.

3 Result and Discussion

To answer RQ1, top-10 strongest collocates were extracted, whose output can be observed in **Table 1**. The top-three collocates are *umat* 'adherents', *kaum* 'group', and *busana* 'outfit'. It can be observed that none of the collocates are negative in terms of their sentiments. The first four strong collocates represent that muslims are the majority. The fifth collocate *non* is a modifier for muslim (non-muslim). This represents diversity of people adhering to other religions (christian, buddhist, hindu, confucious) in Indonesia even though muslims are still the majority.

We present our findings in comparison with other studies, namely Baker et al. [5], Rheinhart [6], and Arifa et al. [7]. Note that their collocation measures are different. Baker et al. [5] used Log Dice statistic, while Rheinhart [6] and Arifa et al. [7] use frequency counts. Regardless, we observe some shared collocates. This is systematically presented in **Table 1** by color.

Collocates with the same color are shared across findings. For example, 'community' is shared with Baker et al. [5] and 'majority' is shared with Rheinhart [6], regardless of the order of the strength of association to muslims. Some of the collocates are unique to our findings, as shown in 'Rohingya' and 'Bukhari'; both are

[1] https://cqpweb.lancs.ac.uk/lccindonesianv3/

proper names. The presence of the first collocates may be driven by Rohingnya refugees news, most are muslims, who took refuge in Indonesia. Bukhari is the name of a muslim Imam whose teachings are practiced in Indonesia. The fact that they do not appear on the remaining findings may be driven by different language, timeline and sources analysed ([5]; [6]) or the paucity of the target corpus [7]).

Table 1. Top-10 collocates: comparison with Baker et al [5], Rheinhart [6] and Arifa et al [7]

No.	Collocates LCC	Dice Coefficient	Collocates Baker	LogDice	Collocates Rhinehart	Freq.	Collocates Arifa et al.	No of Freq.
1	*umat* 'people'	0.08	community	10.4	world	70	*Pemuda* 'youth'	169
2	*kaum* 'group'	0.05	world	9.19	group	36	*Masjid* 'mosque'	150
3	*Busana* 'Outfit'	0.04	woman	9.07	extremist	34	*Remaja* 'teenager'	133
4	*mayoritas* 'majority'	0.03	country	8.74	brotherhood	31	Islam 'Islam'	132
5	*non* 'non'	0.02	leader	8.8	sunni	30	Indonesia 'Indonesia'	83
6	Rohingya 'Rohingya'	0.02	cleric	9.36	countries	2	*Kegiatan* 'activity'	53
7	Bukhari 'Bukhari'	0.02	man	7.5	leaders	26	Muslim 'Muslim'	42
8	*seorang* 'someone'	0.02	group	7.7	leader	25	*Anak* 'child'	35
9	*komunitas* 'community'	0.02	population	8.9	fundamentalist	23	*Bangsa* 'nation'	34
10	*sesama* 'fellow'	0.02	extremist	8.8	majority	22	*Ketua* 'leader'	30

As for RQ2, we identify to what extent the top-10 collocates appear on a yearly basis. Some collocates are very consistent. Collocates such as *umat* and *mayoritas* appear every year. Some other collocates like *komunitas* and *busana* fall beyond top-10 collocates in one year only. We observed that some collocates appear only on certain years. Rohingya, for example, only appears at 2012, 2013, 2015, 2016, and 2017. See more in **Table 2**. Reasons for this may vary. It is possible that texts regarding the topic were absent or less frequent the other years. However, some may question why these collocates are within the top-10 strong collocates (RQ1). The collocate *seorang* 'a (human)' is present in only 2018 and 2022. One that we can verify is that they are also strong collocates even outside the top-10.

Table 2. Partial Collocate Results of Muslim Lema in Indonesian LCC 2023 by Year

Collocates	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
umat 'people'	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
mayoritas 'majority'	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
cendekiawan 'intellectual'	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
kaum 'group'	v	v	-	v	v	v	v	v	v	v	v	v	v	v	v
non 'non'	v	v	v	v	v	v	-	v	v	v	v	v	v	v	v
komunitas 'community'	v	v	-	v	v	v	v	v	v	v	v	v	v	v	v
busana 'outfit'	v	v	v	v	v	v	-	v	v	v	v	v	v	v	v
Rohingya	-	-	-	-	v	v	-	v	v	v	-	-	-	-	-
Bukhari	-	-	-	-	-	v	v	v	v	v	-	-	v	-	-
seorang 'a (human)'	-	-	-	-	-	-	-	-	-	-	v	-	-	-	v

As for **RQ3**, we expanded out collocates search to top-50. Our findings are as follows (1) Ethnic/Nationality Identity (42%) e.g., Uighur, (2) Religion (24%) e.g. *islam* 'Islam', (3) Characterizing/differentiating attributes (20%) e.g., *cendekiawan* 'intellectuals', (4) Culture (10%) e.g., *busana* 'output', and (5) Group/Organization (4%) e.g., *kalangan* 'group'. We can observe that Ethnic/National Entity category is dominant over other categories. Thus, it is very likely that the representations of 'muslim' in muslim-majority and muslim-minority are very different. Interestingly, Arifa et al. (2023) still discovered some conflict related collocates. However, we discussed earlier that Arifa et al.'s corpus is small. Its representativeness is not

comparable to LCC Indonesian 2023 which consists of more than 500 million word tokens with various sources, not just Jawa Pos as Arifa et al. [7] studied.

Table 3. Sample concordance lines of two collocates of 'muslim'

Dalam sebuah hadist yang sangat populer (mutawatir) yang diriwayatkan oleh ***Bukhari***	**Muslim**	*Dan Abu Daud, " khoirunnas anfauhum linnas " artinya , sebaik-baik*
dibuka apa yang ditutupi oleh siapa pun . Aa Gym meminta seluruh ***umat***	**Muslim**	*. Ada 2 hal yang mendasari pelantikan pejabat struktural kali ini .*

4 Conclusion

Our findings suggest the representation of 'muslim' in LCC Indonesian 2023 is different than that of Baker et al. [5] and Rheinhart [6]. The difference in the target language may drive this, as they studied English corpora, and the textual representation. We also highlight the issue of representativeness in the study. The fact that no conflict category collocates were found, which stands in contrast to Arifa et al. [7], is not because there is no conflict-related vocabulary at all in the texts, but because our target corpus is more representative and therefore the conflict words may not be captured as strong collocates.

References

1. Jaszczyk-grzyb, M., Szczepaniak-Kozak, A., & Adamczak-Krysztofowicz, S. (2023). A corpus-assisted critical discourse analysis of hate speech in German and Polish social media posts. *Moderna Sprak, 117(1),* 44-71. https://doi.org/10.58221/mosp.v117i1.11518
2. Lestari, T. E., Anwar, M., & Prihantoro, P. (2024). The Representation of Islam in European Parliament. Sessions: A Corpus Study of 'Europarl 3: German'. *Jurnal Ilmiah Global Education, 5(4),* 2100–2108. https://doi.org/10.55681/jige.v5i4.3504
3. Al-Qattan, A., & Abumiera, R. (2020). Muslims in Contemporary American English 1990-2017: Representation of Muslims in News Media. European Scientific, 16(20), 44-65. http://dx.doi.org?10.19044?esj.2020.v16n20p44
4. Knoblock, N 2020 Silent Majority or Vocal Minority: A Corpus-Assisted Discourse Study of Trump Supporters' Facebook Communication Open Library of Humanities, 6(2): 8, pp. 1–37. https://doi.org/10.16995/olh.507

5. Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A Corpus DrivenAnalysis of Representations Around the Word 'Muslim' in the British Press 1998–2009. *Applied Linguistics, 34(3),* 255–278. https://doi.org/10.1093/applin/ams048
6. Rhinehart, C. (2022). Representation of Muslims in Time Magazine: A Corpus Linguistic Study. *Graduate Theses, Dissertations, and Problem Reports.* https://researchrepository.wvu.edu/etd/11215
7. Arifa, Z., Santi, V. N., & Nadifah, M. (2023). Citra Pemuda Muslim dalam Berita Jawa Pos Online: Analisis Linguistik Korpus. *Jurnal AL-AZHAR INDONESIA SERI HUMANIORA, 8(2),* 123-130. http://dx.doi.org/10.36722/sh.v8i2.1776
8. Prihantoro, & Gillings, M. (2025). The Language of Justice: Examining Courtroom Discourse in an Electoral Conflict. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*. https://doi.org/10.1007/s11196-025-10299-4
9. Prihantoro (2025). A Collocation Analysis of Lemma <lahir> in the Indonesian Translation of the Holy Quran. In: Silberztein, M. (eds) Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NOOJ 2024. Communications in Computer and Information Science, vol 2443. Springer, Cham. https://doi.org/10.1007/978-3-031-89810-5
10. Prihantoro. (2022). A Collocation Analysis of 'energy' in Brown Family Corpus. E3S Web of Conferences, 359, 03012. https://doi.org/10.1051/e3sconf/202235903012