Modeling the Sacred: Considerations when Using Religious Data in Natural Language Processing

Anonymous ACL submission

Abstract

This position paper concerns the use of religious data in Natural Language Processing (NLP), which is of special interest to the Ethics of NLP. Religious texts are expressions of culturally important values, and machine learned models have a propensity to reproduce cultural values encoded in their training data. Furthermore, translations of religious texts are frequently used by NLP researchers when language data is scarce. This repurposes the translations from their original uses and motivations, which often involves attracting new followers. This paper argues that NLP's use of such texts raises considerations that go beyond model biases, including data provenance, cultural contexts, and proselytism. We argue for more consideration of researcher positionality, and of the perspectives of marginalized linguistic and religious communities.

1 Introduction

The Association for Computational Linguistics (ACL) is a secular institution. Its constitution, resolutions and policies make no mention of religion other than forbidding harassment on the basis of religion.[1] Nevertheless the Christian Bible and the Islamic Quran[2] are often used in the scientific and professional activities of ACL, as measured by papers published in the ACL Anthology (see Section 3). Some of the reasons that NLP researchers use the Bible are aptly expressed by Resnik et al. (1999). The Bible is the world's most translated book, with translations in over 2,000 languages, and often multiple translations per language. Furthermore, great care is taken with the translations, so from an NLP perspective data quality is high. It is often easily available in electronic form, and is in the public domain, hence free to use. It has a standard structure which allows parallel alignment verse-by-verse. (And some of these reasons, or similar ones, hold of the Quran too.) For these reasons, as recently as 2006 it was said to be "perhaps surprising that the Bible has not been more widely used as a multilingual corpus by the computational linguistics and information retrieval community" (Chew et al., 2006).

Despite the use of sacred texts in NLP, the ethical considerations around such use do not seem to have received prior attention, other than a brief mention by Mager et al. (2023). This position paper contends that responsible secularism demands engaging with the ethical considerations of the use of sacred texts. We take the position that it is important to contextualise religious texts within not just their religious but also their broader cultural and historical contexts. We argue that it is important to acknowledge the motivations, goals, and positionalities of researchers using religious texts, particularly as it is difficult to have intimate first-hand knowledge of the moral and cultural concerns of a religion beyond one's own (if any).

Section 2 provides relevant background material, including discussing the relationship between academic linguistics and missionary linguistics. In Section 3, we present an empirical study of NLP papers using sacred texts. Section 4 discusses a range of ethical considerations when using sacred texts in NLP. In doing so, we consider a range of approaches to the topic, including ethical theories, Indigenous perspectives,[3] human rights, and the AI principles commonly espoused by institutions. Our goal in doing so is not to evaluate past NLP projects that use religious texts, but rather to encourage more reflecting in and on future work. Based on these considerations, we then make some recommendations for the NLP community in Section 5.

[1] www.aclweb.org, accessed September 2023

[2] This paper follows several style guides in using "Quran", although mentions of the alternate Latinization "Koran" are also considered in the corpus studies we report on.

[3] Following style guides such as (Younging, 2018), this paper capitalizes the first letter of "Indigenous".

Religion	Sacred texts	Est. 2020 population	Proselytizing
Christianity	Bible, New Testament, Old Testament	2,382,750,000	Yes
Islam	Quran (alt. spellings include Koran), Old Testament	1,907,110,000	Yes
Hinduism	Vedas, Upanishads, Puranas	1,161,440,000	No
Buddhism	Tripitaka, Mahayana Sutras, Tibetan Book of the Dead	506,990,000	Yes
Traditional Chinese Religion	Zhuangzi, Tao-te Ching, Daozang	310,000,000	No
Judaism	Talmod, Torah, Tanakh, Old Testament	14,660,000	No

Table 1: Some major world religions and their texts. Population estimates are from the US-based Pew Research Center (www.pewresearch.org), which conducts demographic and other research.

2 Background

2.1 Religion

Precisely defining what constitutes a religion might be notoriously difficult (see e.g., Spiro, 2013; Neville, 2018), and lies beyond the scope of this paper. Typical properties of religions center around giving meaning to existence, and include moral values, spiritual beliefs, theistic beliefs, rituals, stories and mythologies, kinship systems and marriage practices, artistic practices, significant locations, and in some cases a language which plays a special role. These are all closely related to questions concerning values.

Like languages, religions may exhibit regional variation and incorporate local practices, so thinking of them as discrete entities may be somewhat misleading. Some widely cited estimates put the number of worldwide religions at several thousand, although these claims are disputed. What seems more certain is that the imminent extinction of many Indigenous languages will accompany an "impending loss of so many religions and worldviews" (Harrison, 2007, p. 153). Acknowledging these challenges, we nevertheless provide a summary of some of the world's most populous religions in Table 1. Surveying or defining each of these religions is beyond the scope of this paper. Also not included here are the various sects and branches within each religion, nor texts which might be important only to specific branches.

One important distinction is that between proselytizing and non-proselytizing religions. The former attempt to convert new populations, whereas the latter do not. The former are more intricately related to historical practices of colonialism—especially in Africa, the Americas, Asia, and the Pacific—and hence also to neocolonial legacies. Some religions hold that a certain language is privileged for communicating sacred texts to the faithful, while on the other hand Protestant Christianity exemplifies a commitment to communicating in vernacular languages. (Article XXIV of the Articles of Religion of the Anglican Church calls for "such a Tongue as the people understandeth".) Although originally in Ancient Greek, Christian texts have been widely distributed in many European languages since the Protestant Reformation, and the global reach of Christianity is associated with European colonial practices.

2.2 The Academy and Bible Translation

The September 2009 issue of the journal *Language* has a special feature of five articles by anthropologists and linguists concerning the relationships between the US-based Bible translation organization SIL International (SIL) and academic linguistics. In this issue, Dobrin and Good (2009) explore how academic linguists have at times become reliant on, and benefitted from, the technological infrastructures of SIL, in part because creating and maintaining these infrastructures has not been valued by the academy. This "partnership of convenience" causes tensions between differing objectives, and raises questions about what kind of relationships secular research institutions should have with organizations with very different agendas. These practices presage similar ways in which some areas of NLP research have become reliant on Bible translations.

Many linguists and NLP practitioners working on Indigenous languages see their research as addressing issues of human rights and cultural extinction. However, as Dobrin and Good point out, languages which are most endangered are least likely to receive SIL's attention. Handman (2009) draws attention to how SIL ideology separates linguistic identity from religious identity, differing from UNESCO's position that sustaining endangered languages entails sustaining cultural worldviews, knowledge systems, and identity practices. Epps and Ladley (2009) argue that evangelical success entails the displacement or transformation of

Search term/phrase	Results [min, max]
bible	[1920, 3890]]
quran	[291, 547]
new testament	[294, 294]
koran	[131, 248]
old testament	[73, 206]
torah	[25, 153]
talmud	[21, 22]
vedas	[22, 51]
tripitaka	[7, 7]
upanishads	[6, 6]
mahayana sutras	[4, 4]
tanakh	[3, 4]
zhuangzi	[3, 3]
puranas	[3, 3]
tibetan book of the dead	[0, 0]
tao-te ching	[0, 0]
daozang	[0, 0]

Table 2: Number of search results for religious texts in the ACL Anthology on August 10, 2023. Since search result counts are non-deterministic, we report the min and max of 10 searches for each term.

traditional beliefs, often leading to social upheaval, and argues that the academy has a moral interest in supporting local self-determination which is at odds with evangelical agendas.

3 The Use of Sacred Texts in NLP

This section presents empirical data on the use of religious texts in the field of NLP.

3.1 Sacred Texts in the ACL Anthology

In this section we consider four research questions: i) *which* sacred texts are used in NLP?, ii) how does NLP *characterize* those sacred texts?, iii) how does NLP *use* sacred texts?, and iv) are there any *trends* over time? To explore these, we use the Anthology of the Association of Computational Linguistics[4] (henceforth, the ACL Anthology) (Bird et al., 2008), which we consider to represent the research publications of the NLP community.

Regarding *which* sacred texts are used in NLP, the number of ACL Anthology entries for sacred texts are shown in Table 2.[5] Thousands of papers in the ACL Anthology seem to use religious texts. There is a strong bias towards the texts of the Judaism, Christianity, and Islam.

To understand how sacred texts are *characterized* in NLP, we analysed ACL Anthology papers containing the phrases "the Bible is" or "the Quran is". Thematic analysis of the results reveal that descriptions of the Bible mostly emphasize its availability, convenience, size, and multilingual character (see Appendix A). Only infrequently is the religious nature of the Bible mentioned (in stark contrast with how Christian websites describe the Bible). One paper describes the Bible as familiar. NLP papers are more likely to mention religious aspects of the Quran. No ACL anthology papers mention "the Koran is", nor related appositions "Koran, a[n]"; presumably the papers mentioning 'Koran' felt it needed no explanation.

To understand how NLP *uses* sacred texts, we examined the first 100 results (sorting by relevance) on the ACL Anthology for each of the terms 'bible' and 'koran'. We manually looked at each of the 200 search results in order to verify its relevance to our study, and omitted 48 search results for not being relevant (see Appendix B). We manually coded the 88 papers mentioning the Bible for their application domain. The most common application domain was machine translation (23%), while many (18%) introduced a new corpus or lexical resource. Three papers were concerned specifically with Bible translation, and four with literary analysis of the Bible. A variety of other application domains were represented, including various morphosyntactic analysis tasks, and language modeling or pretraining (see Appendix C). 48% of the 88 papers concerned one or more Indigenous languages.[6] The 64 ACL papers mentioning 'koran' were less varied and largely (73%) were concerned with machine translation, typically using verses from the Quran in evaluation datasets (see Appendix D).

To get a sense of *trends* in the use of sacred texts by the NLP community, we examined the year of publication of the 150 of the 152 papers described above which were published between Jan 2004 and July 2023 (just 2 were published prior to 2004). Of these, we see increasing use over time of mentions of the Bible, and especially of mentions of 'Koran' (see Figure 1)[7], including over 60% of papers being published between 2019 and 2023 alone. Of course, the number of papers published in the ACL anthology has also been increasing, for example ACL 2023 accepted about 12 times as many papers

[4] www.aclanthology.org, accessed August 2023.

[5] Non-deterministic result counts seem to be an artefact of the ACL Anthology's use of Google's Programmable Search Engine (Bollmann et al., 2023).

[6] These categorizations were done by the author, taking into account historical and social context, however an ideal approach might engage with language communities to understand whether they consider themselves Indigenous.

[7] Since our 2023 sample was limited to Jan-July, paper counts for 2023 are multiplied by 12/7 when creating sparklines, to make them comparable with previous years.

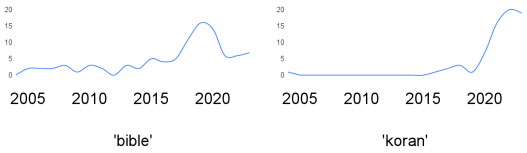

Figure 1: Trend sparklines for counts of papers in the ACL Anthology mentioning 'bible' or 'koran' (see §3).

as ACL 2004 did. We are not necessarily seeing an increase in terms of the fraction of NLP papers using the Bible, and thus there may only be weak evidence of trends *within* NLP research. We do however hypothesize that the absolute counts of papers using religious texts may be a better proxy for the *externalities* (both positive and negative) of NLP research on linguistic and religious communities, and thus relevant to this paper's topic of ethical considerations. There also seems to be a trend towards papers using sacred texts handling very large numbers of languages, with ten papers since 2014 in our sample handling over 500 languages.

3.2 Four Contemporary Case Studies

To complement the broad analysis of the ACL Anthology above, we also report here on some recent noteworthy NLP papers which have attracted acclaim and attention. These illustrate in more detail ways in which the NLP community is encountering and using religious texts.

Our first example is a paper which was uploaded to arXiv, a popular online archive for computer science papers, in May 2023 (Pratap et al., 2023). This work aims to improve Speech Recognition and Speech Synthesis for over a thousand languages. They train their model using translations of the New Testament, as well as audio of readings of those translations, obtained from Faith Comes by Hearing (faithcomesbyhearing.com), goto.bible and bible.com. They also use spoken recordings in many languages, without paired texts, of Bible stories, evangelistic messages, scripture readings, and songs, obtained from Global Recordings Network (globalrecordings.net), whose mission is to communicate "the Good News of Jesus Christ" via a strategy of recording, distribution, and promotion.

Our second example was awarded the ACL Area Chair Award for best Multilingualism and Cross-Lingual NLP paper in July 2023 (ImaniGooghari et al., 2023). This work aims to scale language models to 500 languages. They "crawl or download" data from 150 sources, including religious texts and observe a "higher proportion of religious data" compared to previous comparable work. Parallel verses from Bible translations are used for model training and testing, and performance is reported for Sentence Retrieval from the Bible.

Our third example concerns JW300 (Agić and Vulić, 2019), a dataset of around 100k sentences in each of 300 languages crawled from jw.org, a website run by the US-based Jehovah's Witnesses, a Christian denomination. A majority of the texts come from the Jehovah's Witnesses' magazines *Awake!* and *Watchtower*. Released as a corpus in 2019, JW300 has been cited over 160 times as of December 2023. The African grassroots open-source NLP project Masakhane (masakhane.io) had been using JW300 to train Machine Translation models, until receiving legal advice in 2023 that this was breaching copyright. A subsequent request by Masakhane to the Jehovah's Witnesses for permission to use the data was declined.[8]

Our final recent example concerns the release of MADLAD-400, a new text dataset containing 3T tokens in 419 languages (Kudugunta et al., 2023). It uses 2022 snapshots of the CommonCrawl web crawl (commoncrawl.org) and the paper was uploaded to arXiv in September 2023. Auditing of a preliminary version of the dataset, spanning 498 languages, revealed that for 141 languages there were "significant amounts" of Bible data. Significant amounts of Jehovah's Witnesses data was also found for 37 languages, and of Church of Jesus Christ of the Latter Day Saints (LDS) data for 2 languages. (No Quran data was reported to be found in significant amounts.)

4 Considerations around sacred texts

Having demonstrated above that religious texts have been used in thousands of NLP papers, we now discuss some of the ethical considerations. Our goal is not to critique the works identified in the previous section, but rather to provide a toolbox for assisting critical thinking in the future. In this section we lay out our positions that relevant considerations include not just technological outcomes, but also cultural standpoints, cultural inclusion, and relationships between global and marginalized cul-

[8] See, e.g., https://walledculture.org/a-blatant-no-from-a-copyright-holder-stops-vital-ling or https://www.youtube.com/watch?v=mbkuRZkg1RY.

tures. With this in mind, we now consider the use of sacred texts in NLP from a range of lenses.

4.1 Ethical Theories

Ethical theories are attempts at explaining why actions might be called right or wrong. In this section we consider the topic of sacred texts through just two ethical theories, however our broader point is that if analyses of NLP projects focus narrowly on using a single theory then they risk minimising concerns which other theories bring more focus to.

The term **consequentialism** refers to a family of normative theories which emphasize the importance of considering the consequences of actions, including analysis of risks and benefits. One example of a consequentialist theory is **utilitarianism**, which holds that the best action is one that maximizes wellbeing and minimizes suffering. One challenge with applying these theories in an NLP research context is that in practice research activities are often far removed from applications, and any eventual path between the two can be unknowable or uncertain. Even if the impacts on users of NLP applications can be known, challenges arise in calculating aggregate benefits across disparate stakeholders with different objectives. Perhaps the easiest consequences to reason about are the impacts of NLP research and applications on the lives of the researchers and developers themselves, since research papers and software bring rewards in the forms of measurable kudos, citations, career progression, etc. How can these be weighed against even low-likelihood risks of causing serious offense? If translations of religious texts are purchased from a proselytizing organization, what are the downstream impacts of the organization's reinvesting those payments into their projects?

The term **deontology** refers to a normative theory which posits that there are rules or principles which determine the rightness of wrongness of actions, rather than the consequences. Within an NLP context, this might lead to a focus on the upstream actions such as sourcing of data, rather than downstream actions such as usage of NLP applications. Within the context of religious texts and their translations, this might lead to questions such as: Were the appropriate people involved in the creation of a dataset of sacred texts? Did they have the right roles and relationships from within the perspective of followers of the religion? Do translators of sacred texts have the right specialist linguistic, religious, and and cultural knowledge? Were the translators paid fairly? Were consent practices around voice recordings followed? Generally, was the dataset collected in a manner aligned with best research practices, e.g., as operationalized by research ethics boards? Prabhumoye et al. (2020) discuss the importance of informed consent for deontological approaches to NLP ethics, and community-level consent might be an appropriate lens for thinking about communities of religious practice.

4.2 AI Principles

We use the term **AI Principles** here to refer broadly to the hundreds of sets of principles for responsible and ethical AI that have been released by companies and governments in recent years. These principles might be considered to aspire to be more practical than ethical theories, and are often include technological considerations. Although all unique in their own way, the various sets of principles also have many facets in common (Floridi and Cowls, 2022), such as safety, privacy, bias, fairness, accountability, and transparency. We expect the AI Principles-related concerns to be highly dependent on NLP system functioality and thus on the application domain. Given the frequent use of sacred texts for training and evaluating MT systems, we provide a sketch in Figure 2 of how AI Principles might raise different flavors of considerations.

4.3 Cultural Considerations

Authors of NLP publications do not represent global diversity (Rungta et al., 2022), and hence special consideration needs to be given to cultural standpoints. A useful starting premise might be that ethical consideration of ACL papers should not be biased against nor towards any religions. This might seem to suggest that **Rawls' Veil of Ignorance** could provide a guide (Rawls, 1971): that is, one's consideration of the use of sacred texts within NLP should proceed as if we are each in ignorance of which is our own religion (if any).

Although the Veil of Ignorance might seem attractive at face value, we take the position that our ability in practice to truly avoid being informed by our own cultural backgrounds and affiliations is extremely questionable. We must instead accept that we have cultural standpoints. The **emic/etic distinction** originated in linguistics in the 1950s for describing different standpoints for language research (Mostowlansky and Rota, 2020). *Emic* is

AI Principle	Example possible considerations
Safety	Could the errors made by an MT model trained on sacred texts perpetuate harmful misinformation and hate speech about religiously significant topics?
Privacy	In speech-to-speech translation, could spoken recordings of religious texts (cf. Section 3.2) enable speakers to be identified and their religious beliefs compromised?
Bias and fairness	In what specific ways are religious texts not representative of other domains (cf. Mayhew et al., 2017; Adelani et al., 2021)? Do these biases cause MT models to have unfair impacts?
Accountability	What mechanisms exist for non-English speaking religious communities to report and escalate offensive translations resulting from the use or religious texts? Who is responsible for responding to such reports and mitigating the harms?
Transparency	Do we have accessible evidence that religious communities were consulted when building corpora of parallel religious texts to use for MT?

Figure 2: Non-exhaustive list of possible AI Principles considerations related to using religious texts for MT.

commonly used to describe research on a culture from the perspective of people of that culture. This contrasts with *etic* research, which takes an outsiders perspective. When NLP handles religious texts, we can distinguish research problems and applications which are within the researcher's own religious context, from those applications which impact those having other religious beliefs (e.g., translation for the purpose of proselytizing).

4.3.1 On relativism and the ACL community

One possible objection to the claim that greater consideration of NLP's use of sacred texts is needed is based on an argument of cultural relativism. Such an argument would say that by using sacred texts for an extended period, the ACL community has demonstrated that such practices are judged as acceptable by the norms of the ACL community.

We counter that such an argument would be stronger if the ACL community both had a stronger history of reflexive practices, and was more culturally diverse. Compared to many other disciplines, we find that ACL's interest in ethics to be relatively recent (see Appendix E). The ACL adopts the ACM Code of Ethics,[9] a general code for computing professionals which makes no mention of working with cultural data such as language. Unlike some other disciplines, positionality statements are rare in ACL papers, with "researcher positionality" and "author positionality" each having only a single (and recent) result in the ACL Anthology as of September 2023. We also find that the ACL community does not represent the diversity of the world's local languages and religions; in fact disparities in sources of ACL publications might be increasing (Rungta et al., 2022).

4.4 Power and marginalization

The more different that a culture is from the global cultures which dominate NLP research, the less we should expect the NLP community to understand local concerns. Consideration of marginalized local cultures might bring our focus to questions of how NLP technologies serve to re-arrange power. Which actions are encouraged or discouraged by NLP technologies, and by who? Blodgett has called for NLP researchers to focus their ethical considerations on power relations between technologists and communities (Blodgett et al., 2020). As discussed in Section 3, many NLP papers using the Bible use translations of its texts into Indigenous languages. In this section we give special attention in this section to concerns of local Indigenous communities. Their voices are among the least likely to be represented in ACL's prestigious conferences and journals. And their relationships with global projects are likely to involve large power disparities such that "the asymmetry of power is the cause of domination" (Mager et al., 2023).

4.4.1 Indigenous perspectives

Indigenous concerns regarding **data sovereignty** describe the importance of data governance and Indigenous Cultural and Intellectual Property (ICIP), but also a broader set of concerns regarding practices for projects involving Indigenous data. Surveying the plurality of perspectives (e.g., Carroll et al., 2020; National Health and Medical Re-

[9] https://www.aclweb.org/portal/content/acl-code-ethics

search Council (Australia), 2018; Taiuru, 2021; Janke, 2021; Smith, 2021) is beyond the scope of this paper, however we note they share common themes, which we suggest should be taken into account when working with Indigenous religious texts and/or Indigenous language translations of religious texts. These themes include a) *relationships*, including whether NLP projects display generosity, reciprocity, humility, responsibility, and care, b) *benefits*, including whether NLP projects bring real and equitable benefits to Indigenous communities, while mitigating harms, c) *culture*, including whether the NLP project respects the spirit and integrity of the Indigenous culture, and strengthens the community's connection to their culture, and d) *control*, including whether Indigenous people are in control of their cultural data.

4.4.2 Human rights

As tools of colonizing projects, translations of sacred texts into Indigenous languages have been described as a "well documented example of the non-ethical misuse of translation" (Mager et al., 2023). One consideration for NLP is whether using such translations constitutes complicity with, and promotion of, projects which might violate international human rights to maintain Indigenous cultures.

Kenyan human rights scholar Makua Mutua describes the "basic contradictions" between proselytizing religions and Indigenous cultures (Mutua, 2004). Observing that religion is woven into every aspect of Indigenous social and cultural life, including dances, storytelling, and marriage practices, Mutua argues that such cultures' meeting with proselytizing Christian and Islam faiths amounts to "cultural genocide". In some cases, this characterization seems valid. In other cases, it might be too strong a phrase for the complex realities of syncretic responses to proselytizing cultures. For example, historian Laura Rademaker describes how on Groote Eylandt, in Northern Australia, the Anindilyakwa people reinterpreted Christianity of the missions in their own ways leading to a "hybridisation of cultures" (Rademaker, 2014).

Mutua argues that the right to freedom of religious belief cannot be considered to exist in a "level playing field" in which local cultures can compete with global ones. Rather, the contexts of cultural invasion unfairly privilege global religions, including missionaries making access to education and health services conditional on the "salvation" of "infidels". This echoes arguments by legal scholars that the power and sovereignty dynamics between source and target cultures constitute important factors between proper and improper proselytism (Stahnke, 1999). The Human Rights Committee has acknowledged that the cultural rights protected under Article 27 depend on the ability of a minority group "to maintain its culture, language or religion".[10] As such, Prof. Mutua argues that the (then Draft) UN Declaration on the Rights of Indigenous People appears to prohibit proselytizing by agents external to the the Indigenous culture in order to create space for Indigenous peoples to maintain their cultures amidst external threats.

5 Recommendations

We take the position that the ACL should strive to be aware of risks of harms to religious communities, including dignitive harms, that NLP research may cause or be complicit with. We now suggest five specific actions for the NLP community to mitigate the risks of consequential, procedural, and dignitive harms.

1. Discuss ethical considerations. If one should never speak of religion in polite company, then perhaps ACL forums should be less polite. We suggest that discussions of ethical considerations in NLP papers using religious data should be much more prevalent than they currently are. We propose that Section 4 provides a good starting point for NLP researchers to think about possible concerns, however, building on HCI best practices we also encourage researchers to talk to members of relevant religious communities in order to understand concerns which might lie outside the researchers' domains of lived experience.

2. Consider multiple ethical theories. It has been argued that NLP research often implicitly adopts a utilitarian lens, in the process minimizing other ethical considerations (e.g., Hutchinson et al., 2022). We encourage NLP researchers using sacred texts to contemplate approaching ethical considerations from numerous angles, including not just deontology (discussed in Section 4.1), but also alternatives such as virtue ethics and ethics of care.

3. Delve into domain specifics. We encourage more work on specific AI principle risks of dif-

[10] General Comment No23, UN Doc. ICCPR/C/21/Rev.1/Add.5 (1994), para. 6.2.

ferent NLP domains when using religious texts (cf. Section 4.2). What are the specific ways in which various religious texts are not representative of other domains? What are the model biases that result? Are there specific risks around offensive misinformation on religious topics? Have datasets incorporating sacred texts followed best practices around transparency (e.g. Bender and Friedman, 2018; Gebru et al., 2021; Pushkarna et al., 2022)?

4. Situate NLP work culturally. When should an NLP practitioner tread more carefully? We suggest that when handling sacred data from other cultures and religions there will be more morally consequential risks, since the epistemic and axiological uncertainties are greater, if not the aleatoric ones too. We suggest it is useful to talk of *etic NLP* and *emic NLP* (cf. Section 4.3), according to whether the language technology is for the own linguistic and religious cultures of the NLP practitioner vs. those of others. Linguistic positionality statements have been recommended by researchers who are aware of how different priorities and agendas between researchers and language communities can impact projects (e.g., Rolland et al., 2023; Cormier, 2018), and are exemplified in NLP by Ghosh and Caliskan (2023).[11] Similarly, we suggest that religion positionality statements, for NLP research working with religious data, can also provide a useful signal for the NLP community concerning agendas. For example, hypothetically, would non-Muslim researchers understand the Muslim community's concerns regarding automated mistranslations of the Quran?

5. Attend to concerns of marginalized cultures. Given the NLP community's skew towards the West (Rungta et al., 2022), the values of, and possible harms to, local communities of diverse cultures are not known by most NLP researchers. For Indigenous language NLP projects using translations of sacred texts, we echo calls for more consideration of local colonial contexts, to consider community opinions, and for research to prioritize the needs of Indigenous communities (Bird, 2020; Schwartz, 2022; Alvarado Garcia et al., 2021). Mager et al. (2023) demonstrate one way in which community opinions can be sought regarding NLP projects, and deeper relationships with communities will provide more insights. NLP researchers working with Indigenous languages should become familiar with Indigenous perspectives mentioned in Section 4.4, and with codes of conduct such as that of the Endangered Languages Project.[12] With the imminent extinction of many Indigenous languages and religions, we also suggest there may be a role for NLP to play in documenting and maintaining "first-person accounts of what people once believed in and how they talked to and about their gods" (Harrison, 2007, p, 153).

6 Conclusion

Thousands of NLP research papers have used religious texts, due to their availability, convenience, and multilinguality. However these papers rarely acknowledge the religious and cultural significance of these texts, and the ethical considerations of such use have not previously been explored. This position paper presented a study of the use of religious texts in NLP research, finding that common scenarios include machine translation and dataset creation. We have argued that responsible secularism requires the NLP community to engage with concerns about how such NLP activities might impact religious communities—especially the most marginalized ones—or might be complicit with projects which do. We provided a detailed account of some of the considerations, with a focus on ethical theories, AI principles, cultural considerations, and marginalized communities. We suggested that the field of NLP would benefit from more discussion of the diverse and specific concerns around using religious texts. We proposed that the NLP community engage more with questions of researcher positionality and cultural standpoints, especially with regards to marginalized cultures.

7 Researcher Positionality

I live and work in a secular, multicultural, majority English-speaking, colonized country of the Global North. I am linguistically competent only in English. I grappled while writing this paper with my lack of first-hand experiential understanding of religion, and thus too with my personal role in arguing for more consideration of global and marginalized religious communities.

[11] As of September 2023, the ACL Anthology has no mentions of "language positionality" nor "linguistic positionality".

[12] https://fpcc.ca/wp-content/uploads/2023/02/CodeOfConduct_Web.pdf

Limitations

Any position paper is limited by the experiences of the authors. This is why we include a Researcher Positionality statement.

This paper was unfortunately unable to include lengthy discussion of religious texts in specific NLP domains or specific NLP projects. As the topic of religious texts in NLP has not previously received attention, we took the position that focusing on breadth was most useful. However in Section 5 we do suggest that future work focus more on ethical considerations of using religious texts in specific NLP domains.

We acknowledge that the ACL Anthology used in Section 3 might not be representative of the entire field of NLP, e.g., missing relevant work such as (Chandra and Ranjan, 2022; Bashir et al., 2023), and NLP projects in industry might not be well represented in the ACL Anthology. Another limitation is that we exclude publications in languages other than English. We also acknowledge that the ACL Anthology's non-determinate search result counts are a limitation, seemingly an artefact of the use under the hood of Google's Programmable Search Engine, however reproducing the ACL Anthology's search functionality is beyond the scope of this position paper.

References

David I Adelani, Dana Ruiter, Jesujoba O Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yorùbá-English neural machine translation. *Proceedings of Machine Translation Summit XVIII: Research Track*.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210.

Adriana Alvarado Garcia, Juan F Maestre, Manuhuia Barcham, Marilyn Iriarte, Marisol Wong-Villacres, Oscar A Lemus, Palak Dudani, Pedro Reynolds-Cuéllar, Ruotong Wang, and Teresa Cerratto Pargman. 2021. Decolonial pathways: Our manifesto for a decolonizing agenda in HCI research and design. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Muhammad Huzaifa Bashir, Aqil M Azmi, Haq Nawaz, Wajdi Zaghouani, Mona Diab, Ala Al-Fuqaha, and Junaid Qadir. 2023. Arabic natural language processing for Qur'anic research: A systematic review. *Artificial Intelligence Review*, 56(7):6801–6854.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, Yee Fan Tan, et al. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. Two decades of the acl anthology: Development, impact, and open challenges. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. 2020. The CARE principles for Indigenous data governance. *Data Science Journal*, 19:43–43.

Rohitash Chandra and Mukul Ranjan. 2022. Artificial intelligence for topic modelling in Hindu philosophy: Mapping themes between the upanishads and the bhagavad gita. *Plos one*, 17(9):e0273476.

Peter A Chew, Steve J Verzi, Travis L Bauer, and Jonathan T McClain. 2006. Evaluation of the Bible as a resource for cross-language information retrieval. In *Proceedings of the workshop on multilingual language resources and interoperability*, pages 68–74.

Gail Cormier. 2018. The language variable in educational research: An exploration of researcher positionality, translation, and interpretation. *International Journal of Research & Method in Education*, 41(3):328–341.

Lise M. Dobrin and Jeff Good. 2009. Practical language development: Whose mission? *Language*, 85(3):619–629.

Patience Epps and Herb Ladley. 2009. Syntax, souls, or speakers? on SIL and community language development. *Language*, 85(3):640–646.

Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, pages 535–545.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Courtney Handman. 2009. Language ideologies, endangered-language linguistics, and Christianization. *Language*, 85(3):635–639.

K David Harrison. 2007. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press.

Dirk Hovy, Shannon L Spruit, Margaret Mitchell, Emily M Bender, Michael Strube, and Hanna Wallach. 2017. Proceedings of the first ACL workshop on ethics in natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1876.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Terri Janke. 2021. *True tracks: Indigenous cultural and intellectual property principles for putting self-determination into practice*. University of New South Wales Press.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of Indigenous languages: Giving a voice to the speakers. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545.

Till Mostowlansky and Andrea Rota. 2020. Emic and etic. In Felix Stein, editor, *The Open Encyclopedia of Anthropology*. University of Cambridge.

Makau Mutua. 2004. Proselytism and cultural integrity. In *Facilitating Freedom of Religion or Belief: A Deskbook*, pages 651–668. Brill Nijhoff.

National Health and Medical Research Council (Australia). 2018. *Ethical conduct in research with Aboriginal and Torres Strait Islander Peoples and communities: Guidelines for researchers and stakeholders*. National Health and Medical Research Council.

Robert Cummings Neville. 2018. *Defining religion: Essays in philosophy of religion*. State University of New York Press.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2020. Case study: Deontological ethics in NLP. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

Laura Rademaker. 2014. Language and Australian Aboriginal history; Anindilyakwa and English on Groote Eylandt. *History Australia*, 11(2):222–239.

John Rawls. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33:129–153.

Louise Rolland, Hannah M King, and Pernelle Lorette. 2023. Methodological implications of participant and researcher multilingualism: making language dynamics visible. *Journal of Multilingual and Multicultural Development*, 44(8):645–656.

Mukund Rungta, Janvijay Singh, Saif M Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Lane Schwartz. 2022. *Primum Non Nocere*: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 2.

Linda Tuhiwai Smith. 2021. *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing.

Melford E Spiro. 2013. Religion: Problems of definition and explanation. In *Anthropological approaches to the study of religion*, pages 85–126. Routledge.

Tad Stahnke. 1999. Proselytism and the freedom to change religion in international human rights law. *BYU L. Rev.*, page 251.

Karaitiana Taiuru. 2021. Kaitiakitanga Māori data sovereignty licences. Retrieved from https://www.taiuru.maori.nz/maori-data-sovereignty-licences/.

Gregory Younging. 2018. *Elements of Indigenous style: A guide for writing by and about Indigenous peoples*. Brush Education.

Appendix A: Characterizations of the Bible and the Quran

To select NLP papers for the purposes of examining how they describe the Bible and the Quran, we searched the ACL Anthology for papers mentioning "the bible is" or "the quran is". The Bible is characterized by NLP papers in the ACL Anthology using phrases such as:

- 'available in electronic form', 'most available', 'only available', 'readily available', 'available online' and 'often the only resource available for many languages';
- 'parallel', 'natural source of parallel data', 'ideal source of parallel texts', and 'aligned';
- 'short', 'large', and '[comparatively] small';
- 'massively multilingual', and 'most translated';
- 'central religious text';
- other descriptions including 'narrative of a reasonably straightforward kind', 'representative of core vocabulary', 'familiar'.

The Quran is characterized by NLP papers in the ACL Anthology using phrases such as:

- 'significant religious text', '[believed by Muslims to be] God's word', '[believed by Muslims to be] God's divine words';
- 'fully diacritized';
- 'written in a unique literary style, close to very poetic language'.

Appendix B: Selection of ACL Papers

To select papers to analyze for NLP application domain, we took the first 100 search results on the ACL Anthology for the searches 'bible' and 'koran'. We omitted 12 'bible' search results from our analysis due to either:

- not being in English (2);
- not being research papers (6, e.g., book reviews, invited talks, or proceedings);
- or being duplicates (4)

We omitted 36 'koran' search results from our analysis due to either:

- not being research papers (16);
- being false positive search results (13, e.g., 10 had typos for 'Korean');
- only using the word 'koran' in references (3);
- only using the word 'koran' in footnotes (2);
- only using the word 'koran' in content generated by a model (2).

Appendix C: Application Domains of ACL Papers using the Bible

Application domain	count
Machine translation	20
New corpus or resource	16
Morpho-syntactic tasks other than POS-tagging and dependency parsing	9
POS tagging	6
Language modeling/pretraining	5
Literary analysis of the bible	4
Bible translation	3
Identify text re-use	2
Dependency parsing	2
Other, including sentiment analysis, named entity, CLIR, patronizing language detection	21
Total in selected sample	88
Filtered out (see Appendix B)	12
Total in sample	100

Table 3: Application domains of ACL papers using the Bible.

Appendix D: Application Domains of ACL Papers using the Quran

Application domain	count
Machine translation	47
Text generation	2
Language models	2
NER	2
Other, including hate speech and dialogue act classification	11
Total in selected sample	64
Filtered out (see Appendix B)	36
Total in sample	100

Table 4: Application domains of ACL papers mentioning 'koran'.

Appendix E: A Brief Recent History of Ethics in ACL

- 2017: The first Workshop on Ethics in NLP was held (Hovy et al., 2017).
- 2020: "Ethics and NLP" becomes a possible submission topic for the ACL conference.
- 2022: A Responsible Research Checklist was introduced.
- 2023: Seemingly for the first time, the ACL encourages (but does not require) an ethics statement.