Title	Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic Analysis in the Domain of Theology
Author(s)	Slaughter, Laura; Wang, Wenjie; da Costa, Luis Morgado; Bond, Francis
Citation	Slaughter, L., Wang, W., da Costa, L. M., & Bond, F. (2018). Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic Analysis in the Domain of Theology. Paper presented at The 9th Global WordNet Conference (GWC 2018).
Date	2018
URL	http://hdl.handle.net/10220/44908
Rights	© 2017 The author(s). This is the author created version of a work that has been peer reviewed and accepted for publication by The 9th Global WordNet Conference (GWC 2018). It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_31.pdf].

Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic Analysis in the Domain of Theology

Laura Slaughter
University of Oslo
Oslo, Norway
laurasla@ifi.uio.no

Wenjie Wang, Luis Morgado da Costa♦, Francis Bond
♦Global Asia, Interdisciplinary Graduate School,
School of Humanities,
Nanyang Technological University, Singapore

Abstract

We aim to support digital humanities work related to the study of sacred texts. To do this, we propose to build a cross-lingual wordnet within the domain of theology. We target the Collaborative Interlingual Index (CILI) directly instead of each individual wordnet. The paper presents background for this proposal: (1) an overview of concepts relevant to theology and (2) a summary of the domain-associated issues observed in the Princeton WordNet (PWN). We have found that definitions for concepts in this domain can be too restrictive, inconsistent, and unclear. Necessary synsets are missing, with the PWN being skewed towards Christianity. We argue that tackling problems in a single domain is a better method for improving CILI. By focusing on a single topic rather than a single language, this will result in the proper construction of definitions, romanization/translation of lemmas, and also improvements in use of/creation of a cross-lingual domain hierarchy.

1 Introduction

Sacred texts, including scriptures and exegesis, are the primary source of insight for scholars seeking to understand religious beliefs and practices of past cultures. Scholarly work on ancient manuscripts and papyri has seen an increase in the use of language technologies and automated processing methods. Making texts available in machine-readable formats is a priority within Digital Humanities (DH) projects and new tools are being developed for automated translation and alignment, and also to determine text similarity or patterns of variation. Wordnets play an integral role in improving performance and have been constructed for processing texts in key ancient languages, including Ancient Greek (Bizzoni et al., 2014), Sanskrit (Kulkarni et al., 2010), and Pre-Qin Ancient Chinese (Zhang et al., 2017).

The ideologies of religious traditions are manifested in their sacred texts, with texts being copied, paraphrased, revised, and dispersed over time as a tradition spreads. Various DH projects and tools are developed to process digital editions of aligned multilingual texts. For example, eTrap,[1] the Electronic Text Reuse Acquisition Project has developed a tool called TRACER for assessing text similarities and detecting reuse of parts of texts in later works (i.e. when a portion of text has been appropriated by a later author). It has been used to identify biblical quotes in Swedish literature (Kokkinakis and Malm, 2015) and also to assess similarity of early Christian Coptic texts (Miyagawa et al., 2016, Manuscript submitted for publication.). TRACER makes use of BabelNet[2] which is based on modern language wordnets but does not include ancient languages, leading to less than satisfactory results.

There are now several multi-lingual corpora available in digital formats. For example, the well-known SAT Daizōkyō Text Database (Nagasaki, 2008), a repository making available Buddhist canonical texts in Sanskrit, Tibetan, Chinese, and Japanese, are available for scholars studying the flow of Buddhism from India to China and then into Japan. Projects such as these are expanding their ability to process multi-lingual texts.

Processing of multi-lingual texts in ancient

[1] http://www.etrap.eu
[2] http://babelnet.org

languages is hindered by the inaccuracies of wordnets. An example of this, related to Ancient Greek, is described in Berti et al. (2016). One of the main problems is that wordnets built by bootstrapping from modern languages can be highly erroneous and also miss concepts that are not lexicalized in modern languages (Peters et al., 1998). Bond et al. (2016) proposed the Collaborative Interlingual Index (CILI) as a means to solve just this problem by creating a repository of cross-lingual concepts connecting wordnets.

Future versions of the Open Multilingual Wordnet (OMW) (Bond and Foster, 2013) will use CILI, instead of the current Princeton WordNet (Fellbaum, 1998), to link individual wordnets cross-lingually. This will make it easier to create new concepts that deviate from language specific concept hierarchies, and will invite new ways to link and investigate meaning across languages. The study of sacred texts using DH methods will greatly benefit from CILI, and along with the wordnets for ancient languages will improve results when working with multi-lingual corpora.

This paper discusses the initial steps to organize concepts across wordnets for the case of supporting scholarly work on sacred texts. In the next section, we present results from a survey of the literature in theology that gave us insight into the domain. We then outline specific observations related to coverage, definitions, and structure in Princeton Word Net (PWN). The final section will discuss our proposal that specific domains, such as religion/spiritualism, can be enriched though domain-focused multilingual wordnets.

2 Concepts Related to Theology

The simple question concerning which concepts are most relevant to scholars studying sacred texts is a pertinent one. In an attempt to understand what scholars writing journal articles refer to as a concept within the field of theology, we conducted a search within the theology literature database, ATLA Religion Database® (ATLA RDB®). This database covers the following areas: Bible, archaeology, and antiquities; human culture and society; church history, missions, and ecumenism; pastoral ministry; world religions and religious studies; theology, philosophy, and ethics. There are over 1.8 million records, 667,000+ journal article records, 267,000+ essay records, 607,500+ review records, and 300,500+ book records. The search strategy combined the keyword "concept" in the title and "comparative study" as a subject, making use of ATLA's subject index. The total number of relevant articles retrieved was 116.

The results of this effort are shown in Figure 1. Relevant journal articles were included if they argued that a specific concept was expressed in two or more religious traditions. A wide-range of concepts (e.g. forgiveness) are seen in this literature and also named entities (e.g. God). Terms are usually presented in their original language, and a suitable translation is presented for the English-speaking reader. There were terms provided in diverse languages: Arabic, Hebrew, Sanskrit. Some of the Sanskrit terms are currently standard in the modern English language such as *karma* and *nirvana* although not necessarily with the original meanings.

This review provided us with a starting point. We can begin to think of connections to upper-level ontology categories. Only a few of the concepts from the literature review are related to acts or practices. These are things that a person must actively do and they have endpoints. Forgiving and repentance are acts, or rather processes, with an outcome (for forgiveness, this is alleviation of anger and resentment). Actions can be further divided, some of these are internal processes, for example, deciding and ridding oneself of vengeful thoughts, while others are external actions (including rituals) or deeds having a specific meaning for the doer. Many of the concepts are not actions but experiences that are felt, such as awakening, emptiness, oneness or the experience of humility. Other abstract concepts discussed in the literature are: freedom, justice, and evil. These are non-physical but are observable when acted out.

In this project, we can begin to describe the types of concepts of interest to this domain. We also plan to catalog specific named entities that are found within texts and compare them with those available in existing wordnets, for example, a list of specific places (e.g.

Hell), individuals, supernatural beings, mythical beasts, or objects with magical properties. Some of these entities are no doubt available in wordnets (e.g. Jesus), and we consider how these might be related to other general categories (e.g. prophet).

One question that does come to mind, given the volume of discussion in the literature that resulted in Figure 1, is whether it is at all feasible to connect "theological concepts" in CILI from such diverse languages. This work will be challenging due to the wide variety of concepts expressed in different traditions and how these are interpreted within various cultures. We do expect that it is a feasible task to find cross-lingual equivalents based on prior work on human spirituality. Boyer's (2016) research explains religious concepts in terms of evolved cognitive dispositions and states that these are universal in human minds. He writes "human beings seem disposed to entertain thoughts about non-physically present agents, this includes their thoughts about absent or deceased persons, but also about mythical heroes, fictional characters and a variety of superhuman agents with, usually, counter-intuitive physical capacities but standard mental processes, such as gods, spirits, ancestors, shadows and the like." The anthropologist Donald Brown (2004) produced a compilation of traits that he found were common to all human cultures. From his list, belief in supernatural/religion was one of these traits and there are several others that might be considered elements of common human spirituality, including beliefs about death, divination, empathy, imagery, magic, and moral sentiments.

Nevertheless, given the wide-range of human beliefs, even within a single language there may be problems in defining religious/spiritual concepts, so we can't anticipate that cross-linguistic consensus will naturally emerge. For this reason, it may be the case that some domains are better tackled in domain-focused multilingual wordnets. This would facilitate the proper construction of definitions, romanization/translation of lemmas, and also a better use of/creation of cross-lingual domain hierarchy that could be targeting CILI directly and not each individual wordnet.

3 Examination of Wordnet Synsets

We did two key tasks in order to examine the available synsets in PWN that are relevant to support scholarly work on sacred texts. First, we examined glosses of synsets and possible additions to CILI. Second, we looked at existing categories within PWN: synsets connected to WordNet Domains[3] and hierarchy within PWN.

3.1 PWN Synset Glosses

From Figure 1, we looked for an equivalent English synset for the concept *sunyata,* शून्यता in the Sanskrit WordNet which was translated and discussed as *emptiness* in the literature. *Sunyata,* शून्यता has two available senses in the Sanskrit WordNet, one is connected to the PWN synset for *emptiness* (PWN3.1:`14478672-n` having the gloss *"the state of containing nothing"* and the other is linked to a different translation altogether for *lack, deficiency, or want* (PWN3.1:`14472871-n` *"the state of needing something that is absent or unavailable"*, but neither of these would be accurate according to definitions given by Buddhists who explicitly state that *emptiness* is not defined as *"containing nothing"* (e.g., Suzuki, 2002) and it is never translated in theological terms as *lack, deficiency, want.*

We also looked at the concept of *offering*, which has two noun synsets. One is the contribution (PWN3.1:`13270373-n` *"money contributed to a religious organization"*), and the other the act of giving (PWN3.1:`01041498-n` *"the act of contributing to the funds of a church or charity"*). We found both definitions to be narrower than what we had expected. The former synset's definition restricts offerings to just money, even though other items (such as food) can be offered, and the latter's definition restricts the act to that of only contributing to the church/charity, and only to their funds.

In Hinduism and Buddhism, *daana* (Sanskrit/Pali: dāna दान; *gift-giving* or *generosity*) is the cultivation as well as the practice of generosity and giving. The "gift" in question can be alms; contributions to monasteries and temples, to charity, to the needy; hospitality. Sanskrit has various words

[3]http://wndomains.fbk.eu

- anamnesis
- awakening
- best place
- charisma
- contemplation
- covenant concept
- creation
- devil
- divine action
- divine personhood
- divine providence
- duality
- dyadic nature
- evil
- exile
- forgiveness
- free choice
- freedom
- free will
- justice
- god—creator spirit
- God—supreme being
- Godhead
- good works
- grace—divine grace
- heaven
- hebdomadal
- Holy Spirit
- hospitality
- just war
- justice
- heaven
- Hell
- higher self
- karma
- kingdom, of God
- known by God
- law—natural law
- light
- love— God's love
- loving-kindness[a]
- meditation
- mercy
- messiah concept
- miracle
- new creation
- nirvana
- no-self, also no-I
- nothingness [b]
- paradise (after-life)
- power—supernatural powers
- progressive solemnity
- punitive justice
- reality
- redemption
- repentance
- revelation
- rina[c]
- salvation
- self
- sin
- soul
- spirit
- spiritual perfection
- submission
- time
- ubuntu[d]
- universal savior
- wilderness
- wisdom

[a] "chesed" in Hebrew
[b] translation of "ayin" in Hebrew
[c] debts owed to persons, gods, and ancestors
[d] means personhood, humanness

Figure 1: A Sample of Concepts from ATLA Religion Database®

to describe different types of such offerings. The word *Daana/dāna* (दान) is not found in Sanskrit Wordnet. This would not have been covered by the existing *offering* synsets in PWN. Another related term to *daana*, *Paropakāra* (परोपकार), meaning benevolence or charity, is not found in either the Sanskrit Wordnet nor in PWN. Sanskrit: *Bhiksha* (भिक्षा) is linked to two PWN synsets, *handout* (PWN3.1:01092266-n) and *beggary* (PWN3.1:07202656-n), though it's meaning is closer to an existing PWN synset for *alms* (PWN3.1:01092041-n).

3.2 PWN: Hierarchy and Classifications

The concepts themselves are not always consistent in how they are placed in the wordnet structure and linked to one another. For example, the synset for *Kuan Yin* is defined as *"a female Bodhisattva"*, and *Avalokitesvara* as *"a male Bodhisattva"*. However, only the latter is linked to the synset for *Bodhisattva*.

We also found that related synsets are not always linked to one another. Following up on the previous examples, *Kuan Yin* and *Avalokitesvara* are different forms of the same Bodhisattva (with the former being the East Asian Buddhism variant), but this is not reflected in the relation between the two synsets, nor in their definitions. Also, the two synsets for *Buddha*, one specifically for the historical Buddha (Gautama Buddha), and the other for the concept of a perfectly enlightened being, are also not linked to each other or to *Buddhism*. Synsets for other major figures of various religions — including Jesus Christ and Mohammad — are similarly not linked to their respective religions.

Another concept is that of *spiritual beings*, which has *god* and *satan* as instances, with *angel*, *deity*, *fairy*, etc, as hyponyms. The instances are specific to Christianity (or the Abrahamic religions), and other spiritual beings from other religions are not linked to this synset. We see that the *spiritual being* synset has the hyponym *deity*. Reasonably, the synset for *god* should instead be made a hyponym of *deity*, instead of being an instance of *spiritual being*, albeit that being possible as well. The position of the synset *satan* could likewise be reconsidered. The biblical *Satan*, or the Devil, had been an angel and is now a "fallen angel". Should the *satan* synset be considered under *angel*, or a new hyponymous *fallen angel* concept? This brings to surface the issue of specificness when it comes to the position of the existing synsets in the hierarchy, and how fine-grained such classifications should be made.

In addition to the hierarchy and structure within PWN discussed above, work has been

done to link synsets to domain categories. In Gella et al. (2014), a mapping between WordNet Domains, WordNet topics, and Wikipedia gives us a coarse alignment between WordNet and Wikipedia. The WordNet Domains contain about 200 domain labels that were selected from dictionaries and then structured into a taxonomy based on the Dewey Decimal Classification (Scott, 1998). All of the PWN 1.6 synsets were assigned domains in this project. In all, 2055 synsets are assigned to the domain *Religion*. These do provide a starting point, but the domain labels themselves were assigned roughly based on the conceptual relationships already in PWN and, as we have shown above, there are many issues that must be addressed. The majority of the labeled synsets for Religion are linked to Christian theology, and more specifically to Roman Catholic Christianity. We need to evaluate the accuracy of the labels. *Paradise* (PWN3.1:`05636722-n`) is assigned the label *Christianity* though this is the term generally used in Islam, and the definition given within PWN is not Christian-specific, *"the abode of righteous souls after death"*.

Another question is how to deal with potential relationships between cross-lingual synsets and whether the domain labels assigned reflect useful categories for scholars. Lefebure (1997) equates the Buddhist concept *sunyata* (शून्यता) with *grace*. Looking at *grace* in PWN, there are two synsets with domain labels for Christian theology, (PWN3.1:`14481629-n` *"a state of sanctification by God; the state of one who is under such divine influence"*) and the other is (PWN3.1:`04847946-n` *"the free and unmerited favor or beneficence of God"*). Both *grace* and *sunyata* are states, but the overall discussion of the relationship between these states is complex.

4 Discussion

In the above section, we provided examples for a wide-range of domain-related issues we uncovered in this domain. Synsets are missing from PWN and other wordnets for key concepts in theology (e.g. *emptiness*). We've seen badly formed definitions; those that are too narrow and overly restrictive in addition to having varying details. The hierarchical structure of PWN in this domain is inconsistent. Links to specific instances are missing (e.g. *Kuan Yin*). We have questions concerning the level of granularity and how to deal with fuzzy relationships related to dogma.

We propose a cross-lingual wordnet within the domain of theology; the process of connecting synsets adds to CILI. We believe that creation of such a wordnet will provide a methodology for similar needs in other domains as well as insight that helps correct problems in the single-language wordnets. We start by engaging with experts, scholars who study religious texts, to help with defining religious/spiritual concepts. The proposed method is to make use of the OMWEdit tool (Da Costa and Bond, 2015), a web-based system that is capable of multilingual browsing and editing concepts within OMW. The tool is freely available under an open license and can be used to annotate a corpus. Scholars examining texts in the target languages will be asked to contribute. The goal is to get scholars to provide and annotate parallel texts. These scholars will help us to add new entries to CILI, understand the extent of missing synsets, and make clear the relationships between synsets when it is not possible to identify equivalence.

Wordnets we ultimately wish to examine as part of the proposed work include: Ancient Greek (Bizzoni et al., 2014), Latin (Minozzi, 2009) , Sanskrit (Kulkarni et al., 2010), Pre-Qin Ancient Chinese (Zhang et al., 2017), and Quranic Arabic (AlMaayah et al., 2014). However, we start on a smaller scale, with proposed work based on specific corpora and limiting the number of languages with a focus on depth-first before breadth. Future work will also incorporate on-going results providing CILI definition guidelines (Seppälä, 2015) and the model for diachrous lexical variants from DICOLOD, the Clariah project[4].

Acknowledgments

This research was partially supported by the joint research project on *Multilingual Semantic Analysis* between Fuji Xerox Corporation, Japan and Nanyang Technological University, Singapore.

[4]https://github.com/cltl/clariah-vocab-conversion/blob/master/dicolod-documentation.pdf

References

Manal AlMaayah, Majdi Sawalha, and M Abushariah. 2014. A proposed model for Quranic Arabic WordNet. In *Proc. 2nd Workshop on Lang Resources and Eval for Religious Texts, 31 May*. LRA, pages 9–13.

Monica Berti, Gregory R Crane, Tariq Yousef, Yuri Bizzoni, Federico Boschetti, and Riccardo Del Gratta. 2016. Ancient Greek WordNet meets the dynamic lexicon: The example of the fragments of the Greek historians. In *Proc. 8th Global WordNet Conf*. pages 34–8.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory R Crane. 2014. The making of Ancient Greek WordNet. In *Proc. 9th International Conf on Lang Resources and Eval (LREC'14)*. pages 1140–7.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proc. 51st Annual Meeting of the Assoc for Comp Ling*. Sofia, pages 1352–62.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In *Proc. 8th Global WordNet Conf*. pages 50–7.

Pascal Boyer. 2016. Explaining religious concepts. *Mental Culture: Classical Social Theory and the Cognitive Science of Religion* page 164.

Donald E Brown. 2004. Human universals, human nature and human culture. *Daedalus* 133(4):47–54.

Luis Morgado Da Costa and Francis Bond. 2015. OMWEdit- the integrated open multilingual wordnet editing system. *Proc. ACL-2015, System Demonstrations* pages 73–8.

Christiane Fellbaum. 1998. *WordNet*. MIT Press.

Spandana Gella, Carlo Strapparava, and Vivi Nastase. 2014. Mapping WordNet domains, WordNet topics and Wikipedia categories to generate multilingual domain specific resources. In *Proc. 9th International Conf on Lang Resources and Eval (LREC 2014)*. pages 1117–21.

Dimitrios Kokkinakis and Mats Malm. 2015. Detecting reuse of Biblical quotes in Swedish 19th century fiction using sequence alignment. *Corpus-Based Research in the Humanities (CRH)* pages 79–86.

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. Introducing Sanskrit Wordnet. In *Proc. 5th Global Wordnet Conf (GWC 2010)*. pages 287–294.

Leo D Lefebure. 1997. Awakening and grace: Religious identity in the thought of Masao Abe and Karl Rahner. *CrossCurrents* pages 451–472.

Stefano Minozzi. 2009. The Latin WordNet project. In *Latin Ling Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*. volume 137, pages 707–716.

So Miyagawa, Marco Büchler, and Heike Behlmer. 2016, Manuscript submitted for publication. Computational analysis of text reuse/intertextuality: The example of Shenoute Canon 6. In *Proc. 11th International Congress of Coptic Studies. Orientalia Lovaniensia Analecta.*. Leuven: Peeters.

Kiyonori Nagasaki. 2008. A collaboration system for the philology of the Buddhist study. *Digital Humanities 2008 Book of Abstracts* pages 262–3.

Wim Peters, Piek Vossen, Pedro Díez-Orzas, and Geert Adriens. 1998. Cross-linguistic alignment of wordnets with an inter-lingual-index. *Computers and the Humanities* 32(2/3):221–251.

Mona L Scott. 1998. *Dewey Decimal Classification*. Libraries Unlimited.

Selja Seppälä. 2015. An ontological framework for modeling the contents of definitions. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 21(1):23–50.

Daisetz Teitaro Suzuki. 2002. *Mysticism: Christian and Buddhist*. Courier Corporation.

Yingjie Zhang, Bin Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2017. Pqac-wn: constructing a wordnet for Pre-Qin Ancient Chinese. *Lang Resources and Eval* 51(2):525–45.