Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2833–2840
Marseille, 20-25 June 2022

IndoUKC: a Concept-Centered Indian Multilingual Lexical Resource

Nandu Chandran Nair*, Rajendran S. Velayuthan†, Yamini Chandrashekar*, Gabor Bella*, Fausto Giunchiglia*
*University of Trento, Italy
†Amrita Vishwa Vidyapeetham, India
nandu.chandrannair, yamini.chandrashekar, gabor.bella, fausto.giunchiglia@unitn.it
rajushush@gmail.com

Abstract

We introduce the IndoUKC, a new multilingual lexical database comprised of eighteen Indian languages, with a focus on formally capturing words and word meanings specific to Indian languages and cultures. The IndoUKC reuses content from the existing IndoWordNet resource, while providing a new model for the cross-lingual mapping of lexical meanings that allows for a richer, diversity-aware representation. Accordingly, beyond a thorough syntactic and semantic cleaning, the IndoWordNet lexical content has been thoroughly remodelled in order to allow a more precise expression of language-specific meaning. The resulting database is made available both for browsing through a graphical web interface and for download through the LiveLanguage data catalogue.

Keywords: wordnet, multlingual lexical database, Indian languages, language diversity

1. Introduction

The Indian subcontinent, home to hundreds of languages, presents an impressive linguistic, social, and cultural diversity. This diversity is evidently manifested in the lexicons, with many words and even entire cultural domains lacking precise equivalents outside of India, and often even within the subcontinent. For example, the Malayalam word ഹസ്തസൂത്രം (*hasthasoothram*) represents a type of ornament that only married women wear. This word cannot be found in western lexical resources since this Indian word is not that popular.

Our aim is to give justice to the richness and diversity of Indian lexicography by proposing the *IndoUKC*, a new lexical database for Indian languages. The IndoUKC provides words and cross-lingually mapped word meanings over 18 languages. The distinguishing feature of the IndoUKC is that the lexicons are mapped together in a diversity-preserving manner, meaning that both meaning equivalence and untranslatability—when a word has no equivalent in another language—are explicitly marked. Through a layer of supra-lingual lexical concepts, we have mapped the 18 Indian lexicons to Princeton WordNet and to the lexicons of thousands of other languages.

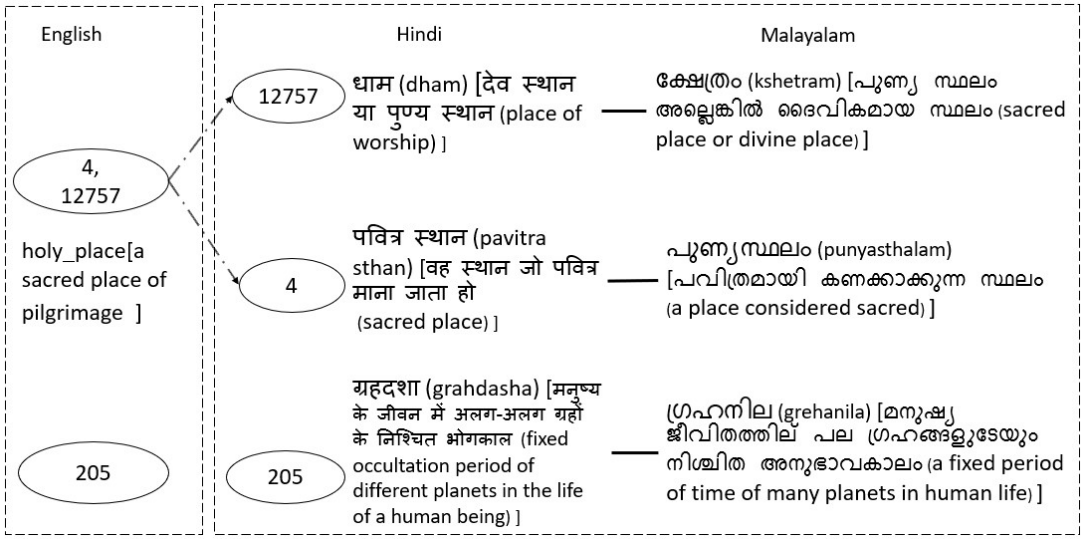
A major data source of IndoUKC is *IndoWordNet* (IWN) (Dash et al., 2017), a multilingual wordnet resource that interconnects 18 lexicons using Hindi as a hub language and that is also partially linked to Princeton WordNet (PWN) (Miller, 1998). IWN is used in cross-lingual computational applications such as machine translation (Chakrabarti and Bhattacharyya, 2004). IWN is a rich source of lexicalisations specific to Indian languages and culture(s). However, IWN does not always represent such word meanings explicitly and formally enough to be exploitable in practice, due to the underspecified correspondences its mapping model is able to express. IWN also contains a certain amount of syntactic noise and semantic (mapping) mistakes.

In order to solve these issues, we have carried out a thorough cleaning and re-mapping of IWN word meanings using the 'diversity-aware' lexical model of the Universal Knowledge Core (UKC) (Giunchiglia et al., 2018; Bella et al., 2022). In particular, we performed (1) a syntactic cleaning of IWN lexical entries, fixing or eliminating noisy data; (2) a major revision and extension of mappings towards English (PWN) synsets, and towards all other languages, via the mapping model of the UKC that is based on a *supra-lingual concept layer*; and (3) a more faithful representation of linguistic diversity, both internally to the Indian subcontinent and externally towards other languages, through an explicit indication of language-specific and thus unmappable (untranslatable) word meanings using *language-specific concepts* and *lexical gaps*.

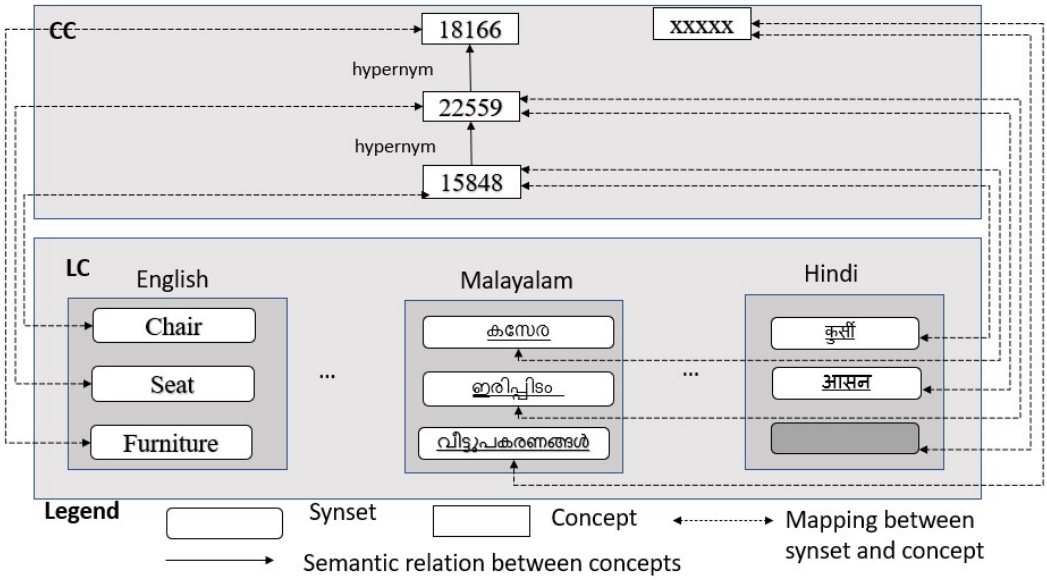
Accordingly, the contributions of this paper are:

- a method for cleaning, re-mapping, and extending IWN in order to obtain interconnected lexicons that properly represent the linguistic diversity of Indian languages;
- the validation of the results by language experts;
- the resulting IndoUKC resource, downloadable from our *LiveLanguage* data catalogue;[1]

[1] `http://www.livelanguage.eu`



(a) IWN



(b) IndoUKC

Figure 1: Diagramatic representation

- the *IndoUKC website*[2] that allows a graphical browsing of IndoUKC data.

We foresee the use of the IndoUKC database both by humans, as a multilingual dictionary that can accurately represent language- and culture-specific words, and by computational applications. It is especially useful for cross-lingual applications, such as cross-lingual transfer or machine translation, for which IndoUKC can indicate both the lack of lexicalisation in the target language (e.g. the Malayalam ഹസ്തസൂത്രം, *hasthasoothram*) and provide a semantically appropriate substitute term, such as *bangle*.

Our paper is organized as follows: Section 2 provides comparison to the state of the art. Section 3 describes the methodology used for producing the IndoUKC. Section 4 evaluates the resource. Section 5 describes the results, including statistics on its current version. Section 6 reflects on followup work.

2. Mapping Models in IndoWordNet and IndoUKC

IWN is a multilingual lexical database with Hindi chosen as a 'hub language', i.e. words of the 18 Indian languages covered are mapped to synsets originally defined for Hindi. English and other Indian languages are connected to Hindi. Furthermore, a small subset of around 60 percent of Hindi synsets is mapped to English PWN synsets (Bhat-

[2] `http://indo.ukc.datascientia.eu`

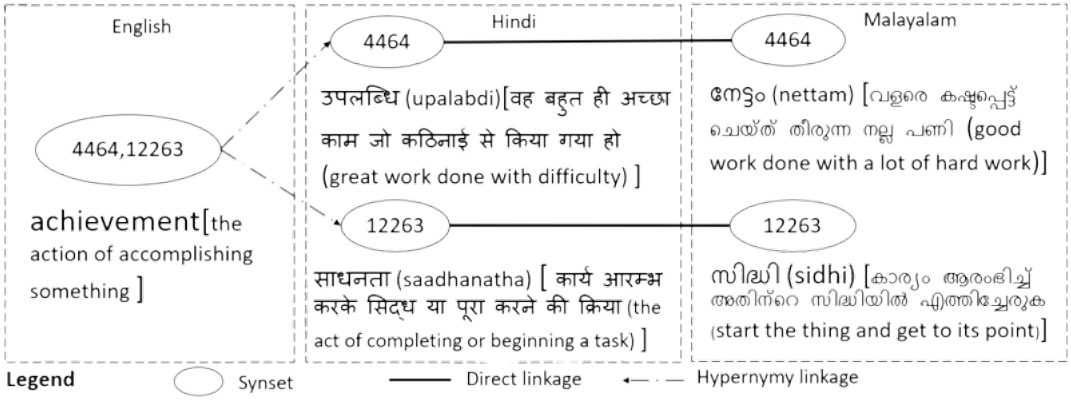


Figure 2: Example of hypernymy linkages in IWN

tacharyya, 2010). The structure of IWN is shown in Figure 1a. Mapping through Hindi is a major advantage of IWN over Western-biased multilingual databases that use English as a hub and, therefore, ignore all lexicalisations that have no English equivalents. While the use of Hindi as hub language is a much more sensible choice for India, it can still lead to biased mappings—by which we understand the incompleteness of mappings that systematically affects certain languages more than others.

The IndoUKC, instead, inherits the mapping model of the UKC that interconnects word meanings through supra-lingual concepts (Figure 1b). As the concept layer of the IndoUKC is not limited to any particular language, bias can be avoided by creating new concepts for yet unmapped meanings.

Beyond using Hindi as a hub, a second novelty of IWN was to use not only equivalence but also cross-lingual hypernymy mappings (Saraswati et al., 2010). Examples of such links are depicted in Figure 2 where the English synset of *achievement* is mapped to two more specific meanings in Hindi and Malayalam. The cross-lingual hypernymy links of IWN are indeed useful to express the existence of words that have no precise equivalents in the hub language. In the IndoUKC, however, we take a different approach by extending the hub—the concept layer—by new concepts that express the more specific meanings. This, in turn, allows the definition of equivalence mappings across both Indian and non-Indian languages through the more specific concepts, something that IWN cannot express due to its limitation to Hindi word meanings in its hub.

The UKC mapping model also allows the definition of *lexical gaps* that formally express lexical untranslatability. Thus, the IndoUKC can map a Malayalam word that has no equivalent in Hindi or English to lexical gaps in these languages, as opposed to confounding untranslatability with resource incompleteness (i.e. omitting the mapping). The explicit indication of lexical gaps is a major device for the diversity-preserving representation of languages and has many computational uses, such as in machine translation (Khishigsuren et al., 2022).

3. Building the IndoUKC

The IndoUKC is the result of addressing correctness and completeness issues in IWN, in particular: syntactic noise, mapping mistakes, and missing mappings. We have implemented a three-step process:

1. cleaning of syntactic noise;
2. a diversity-preserving formalisation and extension of existing cross-lingual mappings;
3. manual validation of the new mappings.

3.1. Syntactic Cleaning

Based on the manual analysis of a sample of IWN data, as well as on our earlier work related to the representation of Malayalam in IWN (Nair et al., 2021), we have gained an understanding of the types of syntactic noise present in IWN, such as empty lemmas or glosses, the use of invalid Unicode characters, or structural inconsistencies. Based on the types of mistakes discovered, we have executed an automated filtering of IWN content. The entries that were identified as erroneous by the filter were then manually analysed and either eliminated or fixed. Table 1 shows the statistics on errors found in IWN. We listed the number of records filtered and cleaned in Table 1.

3.2. Building Diversity-Preserving Mappings

The goal of this step was to transform and extend the underspecified cross-lingual mappings of IWN using the 'diversity-aware' model of the UKC

Table 1: Mistakes identified in IndoWordNet

Sl No	Languages	Total errors	Filtered	Cleaned
1	Assamese	306	0	306
2	Bengali	16498	7	16491
3	Bodo	24	21	3
4	Gujarati	12702	60	12642
5	Hindi	1012	38	974
6	Kannada	129	115	14
7	Kashmiri	14312	33	14279
8	Konkani	7695	110	7585
9	Malayalam	5152	5152	1970
10	Manipuri	44	30	14
11	Marathi	16	0	16
12	Nepali	2174	5	2169
13	Oriya	35267	0	35267
14	Punjabi	18563	11	18552
15	Sanskrit	610	0	610
16	Tamil	590	521	69
17	Telugu	3	2	1
18	Urdu	11416	2	11414

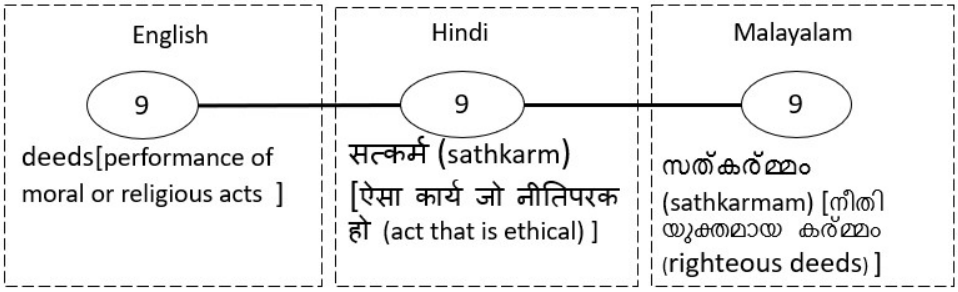


Figure 3: One-to-one (group A) mapping example.

(Giunchiglia et al., 2017). As visible in Figure 2, IWN maps Hindi synsets, on the one hand, to English PWN synsets and, on the other hand, to lexemes from the remaining 17 Indian languages. While the latter mappings, when they exist, are always one-to-one, the English–Hindi mappings can be of three kinds:

- one-to-one mappings (group A): one Hindi synset is mapped to exactly one English PWN synset;
- many-to-one mappings (group B): multiple Hindi synset are mapped to the same English synset through equivalence or hypernymy;
- zero-to-one mappings (group C): the Hindi synset is not mapped to any English synset.

Figure 3 shows an example of a group A (one-to-one) mapping. The English synset of "works, deeds" (*the performance of moral or religious acts*) has one corresponding synset in Hindi, "सत्कर्म" (*sathkarm*) with gloss "ऐसा कार्य जो नीतिपरक हो" (*isa kary jo neethipark ho*) and has one corresponding synset in Malayalam, "സത്കർമ്മം" (*sathkarmam*) with gloss "നീതി യുക്തമായ കർമ്മം " (*neethi yukthamaya karmam*).

Figure 4 shows an example of a many-to-one (group B) mapping. The English synset of "achievement" (*the action of accomplishing something*) is mapped to two narrower synsets in Hindi and Malayalam. The first one with ID 4464 is the Hindi synset of "उपलब्धि" (*upalabdhi*) which means "great work done with difficulty" and the Malayalam synset of "നേട്ടം" (*nettam*) which means "good work done with a lot of hard work". The second one with ID 12263 is the Hindi synset of "साधनता" (*sathantha*) which means "the act of completing or beginning a task" and the Malayalam synset of "സിദ്ധി" (*sidhi*) which means "start the thing and get to its point".

Figure 5, finally, shows an example of a nonexistent (group C) mapping. The Hindi word "आंगनबाड़ी" (*anganbadi*) is Indian terminology for a location to look after children; it differs from "daycare" in that it emerged as a result of an Indian government mission. IWN provides no English mapping

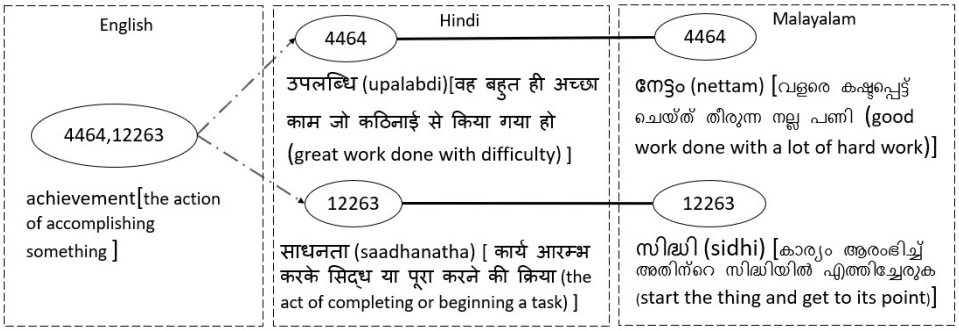


Figure 4: Many-to-one (group B) mapping example

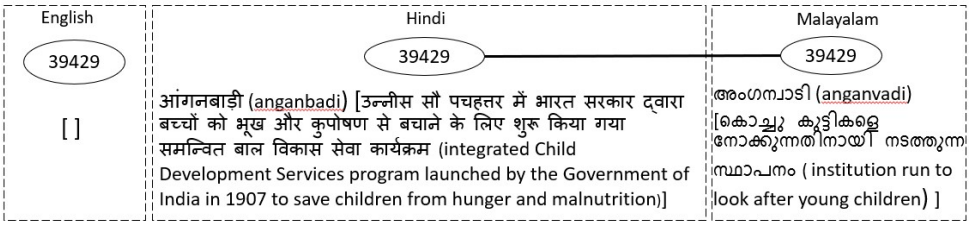


Figure 5: Zero-to-one (group C) mapping example.

for it, nor for the corresponding Malayalam word "അംഗന്വാടി" (*anganvadi*).

Our goal was to convert these mappings into diversity-preserving mappings using the UKC model (described in the previous section), in the following way:

- if there is an equivalence mapping between an English and a Hindi synset which, in turn, is mapped to a lexical entry in, say, Malayalam, then these mappings are converted into two equivalence mappings to the UKC concept that corresponds to the English PWN synset;
- if there is a hypernymy mapping between an English and a Hindi synset, then a manual analysis is performed, which either finds an existing concept equivalent to the Hindi synset or, if such a concept does not exist, creates a new concept.

The operations above cover all one-to-one (group A) and hypernymy (group B) mappings, converting them into equivalence mappings to existing or new concepts. We plan to cover group C (so far unmapped) Hindi synsets in future work.

Based on our early work on group A equivalence mappings (Nair et al., 2019), the aforementioned operations were performed manually by Tamil and Malayalam language experts who also had knowledge of Hindi. Subsequently, the new mappings defined for Hindi, Tamil, and Malayalam were automatically extended to the remaining 15 Indian languages. This step was simple to automate as the IWN mappings between Hindi and the remaining Indian languages always represent equivalence.

3.3. Validation

We used sample sets from the merged resource to validate the mappings between the languages. Our validation consisted of two steps: a review of mappings for group A synsets, followed by an assessment of mappings for group B synsets.

- Validation of group A synsets: a file containing 8325 synsets in English and Malayalam was created. Glosses and synsets were examined in order to confirm whether their mapping was executed correctly and by asking whether a concept from English is tied to the corresponding concept from Malayalam by a relationship of equivalence.
- validation of group B synsets: A file with 6190 mappings in English and Malayalam was prepared, and another with 5666 mappings in English and Tamil. Validators were requested to isolate the records that were appropriately mapped.

4. Evaluation

The goal of evaluation was to check whether the re-mappings of English–Hindi correspondences, done based on Malayalam and Tamil words and glosses and then automatically extrapolated over all 18 languages, were correctly mapping words also from languages other than Malayalam, Tamil, and Hindi.

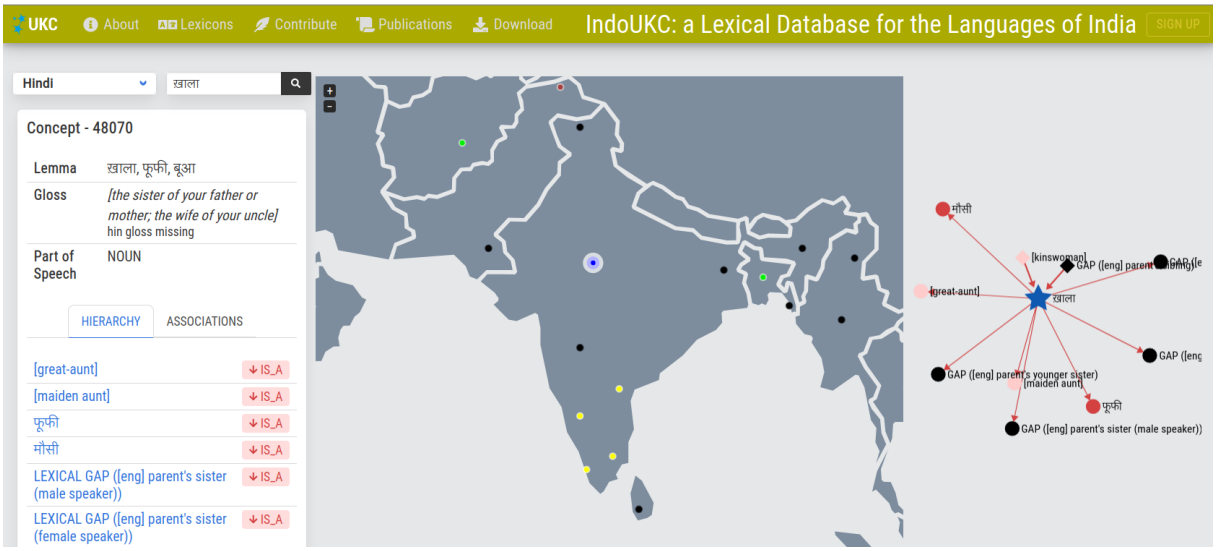


Figure 6: Screenshot from the IndoUKC website, showing the lexicalisation of the concept of *aunt*.

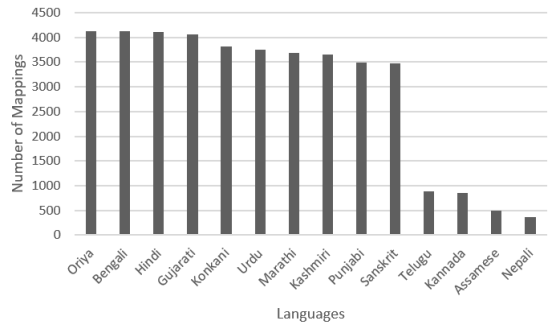


Figure 7: Number of mappings per language in the IndoUKC

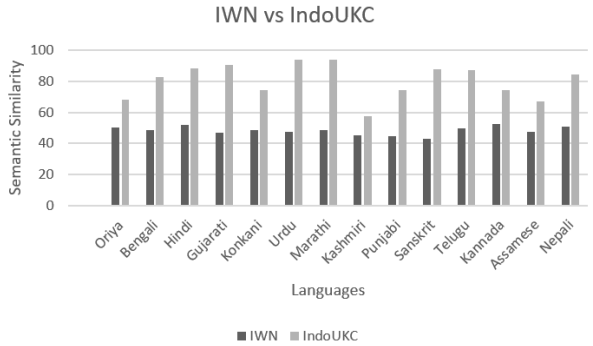


Figure 8: Mappings comparison : IWN vs IndoUKC

Evaluation was done over a sample of 77 noun concepts and 14 languages. We selected concepts tied to mappings that are common across all languages and that are not unique to Indian languages, in order to avoid bias within the evaluation. As shown in Figure 7, the number of original IWN mappings varies greatly among the 18 languages.

We used a measure of semantic similarity between the glosses of source and target synsets in order to evaluate the correctness of mappings. A pre-trained machine learning model was used in order to compute semantic alignment and cosine similarity. There are a number of pre-trained models available for similarity checking. Multilingual BERT(mBERT)(Pires et al., 2019), XLM-RoBERTa(Conneau et al., 2019) and Sentence transformers(Reimers and Gurevych, 2019) are only few of them. Multilingual BERT(mBERT) and XLM-RoBERTa have been known to produce less than ideal sentence representations when deployed out-of-box.They would additionally pose the problem of coming with non- aligned vector spaces across languages. As a result, sentences with the same semantic content, expressed with different languages, would be mapped to different vector spaces. As part of our work we used sentence transformers: a Python framework for state of the art sentence and text embeddings that can be used to compute sentence/text embeddings for more than 100 languages. These embeddings can be compared through cosine similarity, allowing for the identification of sentences with similar meaning. The framework is based on PyTorch(Paszke et al., 2019), Transformers and offers a large collection of pre-trained models tuned for various tasks. We have used the pre-trained multilingual model *stsb-xlm-r-multilingual* and aligned vector spaces allowing for similar inputs across languages to be mapped close within the same vector space. XLM-R supports 100 languages including 13 Indian languages, and is as such able to handle linguistic inputs without the need to specify what the input

language is upfront. The model produces similar embeddings as the bert-base-nli-stsb-mean-token model.

We calculated semantic similarity for English-language and Indian-language glosses, converted into vectors using BERT sentence transformers. We clustered semantic scores depending on each language's overall average—for example, the average semantic score for mappings that involve English–Assamese is 0.4. We chose a minimum 0.4 threshold, following work by (Khodak et al., 2017) who use a machine-readable dictionary to create WordNet data automatically. We classify sets of synsets with cosine similarity scores lower than 0.4 as requiring human verification based on our classification. Hence we considered mappings with a greater than 0.4 as correct association between English and Assamese. Figure 8 shows a semantic similarity comparison between IWN and IndoUKC. Our evaluation showed that human judgment helped mappings of other languages, specifically those with a high number of mappings.

5. Results, Statistics, and Discussion

A new concept-oriented resource of Indian languages was created by using the UKC. The one-to-one and many-to-one synsets from IWN were imported into the IndoUKC based on the validation conducted. Statistics for the current version of the IndoUKC are shown in Table 2.

Table 2 shows the conceptual mappings between IWN and PWN using the UKC based on the three groups. These tables show the three groups of synsets for all Indian languages. An average of 9K concepts are available in each language in IndoUKC. The concepts, which are bolded in Table 2, are available in the current version of IndoUKC, and in the next version remaining concepts (shown in italics in Table 2) will be imported.

The IndoUKC website[3] (Figure 6 shows a screenshot) lets users browse the lexical data through multiple modalities: by word lookup in any of the Indian languages or English (left), by browsing languages on the map (middle), or moving through the interactive concept graph (right). The website also provides download links towards the LiveLanguage data catalogue,[4] where the IndoUKC contents can be downloaded either as individual per-language lexicons or as a multilingual interconnected lexical resource with a user-selected Indian language (or English) acting as a hub.

6. Conclusions and Future Work

This paper described the generation of the IndoUKC multilingual lexical database that provides more precise and more complete cross-lingual mappings for lexical data originally developed for the IndoWordNet resource. The use of supra-lingual concepts instead of language-specific synsets allowed us to represent Indian language- and culture-specific word meanings and untranslatability in a more precise manner.

IndoUKC is an ongoing project. As immediate future work, we will address the coverage of so far unrepresented zero-to-one mappings, that will become lexical gaps in the IndoUKC. The completeness of the resource will also be extended by new words and new languages retrieved and mapped from Wiktionary, by new relationships such as cognates retrieved from (Batsuren et al., 2022), as well as by culture-specific concepts, such as from the kinship domain which demonstrates remarkable lexical diversity and for which exploitable data has been recently published (Khishigsuren et al., 2022), but also by concepts from the *Bhagavad Gita*, among others in order to contribute to the long-term preservation of the Sanskrit language.

7. Acknowledgements

The research has received funding from the "DELPhi - DiscovEring Life Patterns" project funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2017 – DD n. 1062 del 31.05.2019.

We have to thank Maria-Chiara Giangregorio for proofreading the article. We gratefully acknowledge Khuyagbaatar Batsuren's early contributions to the UKC.

8. Bibliographical References

Chakrabarti, D. and Bhattacharyya, P. (2004). Creation of english and hindi verb hierarchies and their application to hindi wordnet building and english-hindi mt. In *Proceedings of the Second Global Wordnet Conference, Brno, Czech Republic*. Citeseer.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Giunchiglia, F., Batsuren, K., and Bella, G. (2017). Understanding and exploiting language diversity. In *IJCAI*, pages 4009–4017.

Khodak, M., Risteski, A., Fellbaum, C., and Arora, S. (2017). Extending and improving wordnet via unsupervised word embeddings. *arXiv preprint arXiv:1705.00217*.

Nair, N. C., Velayuthan, R. S., and Batsuren, K. (2019). Aligning the indowordnet with the princeton wordnet. In *Proceedings of the 3rd in-*

[3] `http://indo.ukc.datascientia.eu`

[4] `http://www.livelanguage.eu`

Table 2: Conceptual mappings of IWN using the UKC

		IWN	UKC				IWN	UKC	
Language	**Group**	**synsets**	**concepts**	**gaps**	**Language**	**Group**	**synsets**	**concepts**	**gaps**
Assamese	A	7314	7314	0	Manipuri	A	7328	7328	0
	B	2646	1779	867		B	2855	1809	1046
	total	9960	**9093**	*867*		total	16351	**9137**	*1046*
Bengali	A	9428	9428	0	Marathi	A	8691	8691	0
	B	10020	1691	8329		B	11265	1688	9577
	total	19448	**11119**	*8329*		total	32721	**10379**	*9577*
Bodo	A	7346	7346	0	Nepali	A	6218	6218	0
	B	2838	1904	934		B	1505	875	630
	total	10184	**9250**	*934*		total	11713	**7093**	*630*
Gujarati	A	9342	9342	0	Oriya	A	9281	9281	0
	B	9899	1762	8137		B	9844	1646	8198
	total	19241	**11104**	*8137*		total	35284	**10927**	*8198*
Hindi	A	11283	11283	0	Punjabi	A	9123	9123	0
	B	13416	454	12692		B	8548	1641	6907
	total	24699	**11737**	*12692*		total	32364	**10764**	*6907*
Kannada	A	7880	7880	0	Sanskrit	A	7213	7213	0
	B	4661	2076	2585		B	22481	1497	20984
	total	12541	**9956**	*2585*		total	38070	**8710**	*20984*
Kashmiri	A	8544	8544	0	Tamil	A	9745	9745	0
	B	8360	1879	6481		B	7123	139	6984
	total	16904	**10423**	*6481*		total	25419	**9884**	*6984*
Konkani	A	9218	9218	0	Telugu	A	7633	7633	0
	B	9413	1720	7693		B	4303	2104	2199
	total	18631	**10938**	*7693*		total	21091	**9737**	*2199*
Malayalam	A	9719	9719	0	Urdu	A	9205	9205	0
	B	10127	1465	8662		B	8949	1659	7290
	total	19846	**11184**	*8662*		total	34280	**10864**	*10864*

ternational conference on natural language and speech processing, pages 9–16.

Nair, N. C., Giangregorio, M.-c., and Giunchiglia, F. (2021). Is this enough?-evaluation of malayalam wordnet. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 100–108.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saraswati, J., Shukla, R., Goyal, R. P., and Bhattacharyya, P. (2010). Hindi to english wordnet linkage: Challenges and solutions. In *Proceedings of 3rd IndoWordNet Workshop, International Conference on Natural Language Processing 2010 (ICON 2010)*.

9. Language Resource References

Batsuren, K., Bella, G., and Giunchiglia, F. (2022). A large and evolving cognate database. *Language Resources and Evaluation*, 56(1):165–189.

Bella, G., Byambadorj, E., Chandrashekar, Y., Batsuren, K., Cheema, D. A., and Giunchiglia, F. (2022). Language diversity: Visible to humans, exploitable by machines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*.

Bhattacharyya, P. (2010). Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Dash, N. S., Bhattacharyya, P., and Pawar, J. D. (2017). *The WordNet in Indian Languages*. Springer.

Giunchiglia, F., Batsuren, K., and Freihat, A. A. (2018). One world–seven thousand languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018*.

Khishigsuren, T., Bella, G., Batsuren, K., Freihat, A. A., Chandran Nair, N., Ganbold, A., Khalilia, H., Chandrashekar, Y., and Giunchiglia, F. (2022). Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. In *LREC*.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.