Check for
updates

Representing interlingual meaning in lexical databases

Fausto Giunchiglia[1,2] · Gábor Bella[1] · Nandu C. Nair[1] · Yang Chi[2] · Hao Xu[2]

Abstract

In today's multilingual lexical databases, the majority of the world's languages are under-represented. Beyond a mere issue of resource incompleteness, we show that existing lexical databases have structural limitations that result in a reduced expressivity on culturally-specific words and in mapping them across languages. In particular, the lexical meaning space of dominant languages, such as English, is represented more accurately while linguistically or culturally diverse languages are mapped in an approximate manner. Our paper assesses state-of-the-art multilingual lexical databases and evaluates their strengths and limitations with respect to their expressivity on lexical phenomena of linguistic diversity.

Keywords Language diversity · Lexical semantics · Multilingual lexical database · Untranslatability

1 Introduction

According to the Ethnologue (Eberhard et al. 2022), there are around seven thousand languages actively spoken in the world today. Despite the immense value—cultural, communicational, economic, etc.—embedded in languages and dialects, whether living or ancient, most computational resources on language have so far focused on a small subset of them, namely those spoken in the richest parts of the world (Joseph et al. 2010). Suggestive studies from Kornai (2013) and Oxford (2015) articulate how digitally less favoured

✉ Fausto Giunchiglia
   fausto.giunchiglia@unitn.it

✉ Gábor Bella
   gabor.bella@unitn.it

   Nandu C. Nair
   nandu.chandrannair@unitn.it

   Yang Chi
   yangchi19@mails.jlu.edu.cn

   Hao Xu
   xuhao@jlu.edu.cn

[1] Department of Information Engineering and Computer Science, University of Trento, via Sommarive, 5, Trento 38123, Italy

[2] College of Computer Science and Technology, Jilin University, Changchun, China

populations suffer from what is called the *Digital Language Divide*, in terms of linguistic and cultural impoverishment. In particular, beyond single-language lexical resources, *multilingual lexical databases* (MLDB) play a pivotal role in language technologies such as cross-lingual word sense disambiguation, machine translation, or multilingual language models. They are also crucial for endangered and minority languages: for putting them in relation with all the world's languages, as reference material for language learners, and as knowledge-driven technology that complements corpus-based approaches in the absence of large corpora.

The goal of this paper is to draw upon these needs and to assess the state of the art in the development of MLDBs. We consider this survey a first step to drive future efforts. The key issue on which we concentrate is that no two vocabularies represent the world in exactly the same way, due to the pervasiveness of diversity in language, culture, and in how reality is perceived differently around the world. MLDBs need to capture these differences in expressivity (Giunchiglia et al. 2017, 2018) and deal with *untranslatability* and cross-lingual shifts of meaning (Catford 1978). A failure to represent the linguistically or culturally specific elements of the vocabulary of a language may lead to a *loss of function* (Kornai 2013) and to an imposed uniformization with the world's dominant languages (Bella et al. 2022a). Our paper has two main contributions:

- a qualitative analysis of state-of-the-art MLDBs, reviewed according to four criteria that together enable an unbiased and diversity-aware representation of interlingual meaning; and
- a complementary, quantitative evaluation of interlingual representation ability of these MLDBs over a corpus of about two thousand gold-standard interlingual mappings from linguistically and culturally diverse lexical fields.

Our analysis makes evident the pervasiveness of lexical untranslatability—the impossibility to find suitable concise translations for a word in another language—and the lack of computational resources that provide such evidence. A second take-home message is that representing interlingual meaning is, before anything else, a problem of lexico-semantic knowledge *structure*: the lexical model underlying a MLDB intrinsically constrains its capability to represent lexical diversity. To scale to all the world's languages, the model needs to be powerful enough to capture, at the very least, the interlingual correspondences used in traditional lexicography: *equivalence* when words "have the same meaning" for practical purposes, *broader–narrower* relationships, and in case of untranslatability, indicating the presence of a *lexical gap* as well as suitable broader terms as alternatives.

The paper is organized as follows. Section 2 provides the theoretical background. Section 3 presents the mapping models of five state-of-the-art exemplary MLDBs. Section 4 provides a quantitative evaluation of a set of relevant reference resources, as well as a comparison of their mapping models. Finally, Sect. 5 provides the conclusion. Throughout the paper we will use the example of family relationships—well known to be expressed in diverse manners across languages (Khishigsuren et al. 2022a)—and in particular the notion of *cousin*, in nine languages: English, French, Italian, Chinese, Hindi, Tamil, Malayalam, Hungarian, and Mongolian.[1]

---

[1] The English *cousin* does not have a precise equivalent in six out of the eight other languages.

Table 1 Examples for interlingual mapping types, used as test cases in the paper, in Malayalam, Tamil, Chinese, and English

Type	Source meaning	Target meaning
Equivalence	MALAYALAM അനിയത്തി: younger daughter of father's sister	TAMIL தங்கை: younger daughter of father's sister
Untranslatability	CHINESE 堂妹: younger female patrilineal cousin	ENGLISH *lexical gap*
Hyponymy	ENGLISH *cousin:* the child of your aunt or uncle	CHINESE 堂妹: younger female patrilineal cousin

2 Cross-lingual lexical mappings

Lexical equivalence is understood by linguists as a complex and multidimensional problem, ranging from multiple coexisting forms of meaning equivalence (Adamska-Sałaciak 2010) to *untranslatability* (Catford 1978) (see Table 1 for examples). While the latter phenomenon, i.e. the absence of certain lexical mappings, cannot be entirely explained through systematic principles (Lehrer 1970), differences from one language to another are often due to diversity in culture or the reality perceived. Some examples are: the lack of vocabulary for sailing in Mongolian, the language of a landlocked country, the Italian word *malga* meaning a kind of mountain restaurant, the Scottish Gaelic *onfhadh* meaning *the raging sound of the sea*, or the rich East Asian vocabulary on the various forms of *rice* as grain and as food.

At the same time, traditional bilingual dictionaries remain pragmatically-built and practice-oriented tools for the general public, typically lacking a fine-grained and theoretically precise modelling of the cross-lingual mapping of meaning (ten Hacken 2016). The relationships provided by dictionaries usually imply a quasi-equivalence of word meanings or, more rarely, a broader target meaning if the target language does not have a close enough word sense. Some dictionaries also indicate *lexical gaps*, i.e. where the target language does not lexicalize the meaning of the source word, as free-text definitions. Furthermore, bilingual dictionaries have always been designed to be asymmetric, clearly defining the source and the target language, and the reverse counterpart is never constructed by the mere inversion of its entries. This is due to translation, even when applied to individual word senses, being by nature asymmetric and intransitive (Adamska-Sałaciak 2010). In the context of MLDBs, however, the principle of asymmetry is never respected in practice, for reasons of scalability: if a MLDB supports n languages then mappings would need to be defined for $n(n - 1)$ language pairs. In order to reduce the number of mappings needed, all MLDBs rely on a *hub* (or *pivot*) meaning representation to which all lexicons are mapped. The possibility of a hub meaning c, however, is based on the simplifying assumption that the mapping of word meanings is an equivalence relation that, by definition, is symmetric and transitive:

$$m_a \leftrightarrow c \leftrightarrow m_b \Rightarrow m_a \leftrightarrow m_b.$$

For this reason, MLDBs tend to rely mostly on equivalence mappings and, instead, express broader–narrower relationships either within their hub or within language-specific lexicons.

The observation above motivates our goal of comparing the cross-lingual semantic expressivity of MLDBs. The first and fundamental evaluation criterion relates to lexical

concepts: it is the ability of the MLDB to represent language-specific lexical meaning. When the hub meaning space of an MLDB is limited to that of a particular language (such as English), it means that the entire database is biased towards that language, as certain lexicons cannot be represented with the same level of detail as others. Beyond the space of meanings, we also evaluate interlingual mapping ability, namely the semantic expressivity of interlingual relations. These should be able to represent interlingual meaning equivalence, but also non-equivalent correspondences and untranslatability, as illustrated in Table 1.

Accordingly, we are going to compare MLDBs with respect to the four criteria below:

1. *Unbiased lexical meaning space* whether the MLDB can represent language-specific lexical concepts for any of the languages it covers, or it is fixed and bound to the meanings from one specific language.
2. *Interlingual equivalence relation* whether the MLDB can express concept equivalence for any language pair (among the languages supported).
3. *Interlingual hypernymy relation* whether the MLDB can express broader–narrower relationships for any language pair (among the languages supported).
4. *Untranslatability relation* whether the MLDB represents lexical gaps as a way explicitly to indicate untranslatability for any language pair (among the languages supported), distinguishing it from the mere absence of a mapping that implies lexicon incompleteness (Bentivogli and Pianta 2000).

Mapping relations beyond equivalence have major uses in cross-lingual applications. For example, a machine translation (MT) system translating the English sentence "This rice is tasty" into Swahili (but also Japanese, Hindi, etc.) can be informed by an MLDB of the fact that Swahili has no equivalent word for *rice* (untranslatability); instead, it has the more specific words *mchele*, meaning *uncooked rice*, and *wali*, *cooked rice* (hyponymy). This knowledge helps the MT system select the best translation depending on the context, *wali*, and avoid the incorrect *mchele* that leads to a translation with the unintended meaning "this raw rice is tasty". An MLDB that does not distinguish untranslatability from lexicon incompleteness—where an equivalence mapping from *rice* to Swahili is is simply missing—will not be able to inform the MT system of the difficulty within the sentence, and a purely corpus-statistics-based approach may lead to erroneous translation, even in state-of-the-art systems such as Google Translate.[2] To our knowledge, the three kinds of interlingual relationships covered by our criteria are on a par with interlingual mappings provided in the best traditional bilingual dictionaries. While in principle we could consider other types of associative cross-lingual relations, such as etymology or cognacy, most MLDBs reviewed in this paper do not contain such information and thus they would not be useful for purposes of comparison.

Throughout the paper we will use the running example of family relationships—well known to be expressed in diverse manners across languages—and in particular the notion of being the *cousin* of somebody, in nine languages: English, French, Italian, Chinese, Hindi, Tamil, Malayalam, Hungarian, and Mongolian. The English *cousin* does not have a precise equivalent in six out of the eight other languages. Instead, they lexicalize more specific concepts among the no less than 63 combinations of *the elder–younger son-daughter*

---

[2] Google translates the sentence above into the Swahili "Mchele huu ni kitamu." Such semantic mistakes are frequent in all major translator tools as of today.

of my father's–mother's elder–younger brother–sister. Thus, in French and Italian, distinct words (inflections) exist to represent the female cousin (*cousin/cousine* and *cugino/cugina*). In Chinese, eight words express the elder–younger son–daughter of your mother's–father's sibling (表姐; 表妹; 表哥; 表弟; 堂姐; 堂妹; 堂兄; 堂弟). Hindi also uses eight distinct words, yet they are not equivalent to the Chinese ones: they express the son–daughter of your mother's–father's brother–sister (फुफेरा भाई; चचेरा भाई; ममेरा भाई; मौसेरा भाई; चचेरी बहन; फुफेरा बहिन; मौसेरा बहिन; ममेरा बहिन). Malayalam and Tamil, finally, each have no less than 16 distinct words to express the elder–younger son–daughter of your mother's–father's brother–sister. Examples such as these cannot be ignored as corner cases. In many societies (such as in Southern India) it is a requisite of appropriate communication to express family relations precisely, and fuzziness is culturally not acceptable. Translators, whether human or AI-based, therefore need to deal with such cases in a correct and coherent manner. While translating any of the specific Chinese, Hindi, or Malayalam words into the more general *cousin* is formally correct (even though information is lost), in the reverse direction a non-semantically-motivated (random or corpus-frequency-based) selection among candidate meanings is likely to inject unintended meaning.

3 Qualitative analysis

Several past and ongoing efforts exist for building lexical resources, with different underlying motivations, solutions, and sizes (Gurevych et al. 2016). Among these, our paper addresses resources that:

- are *multilingual*, as the focus of our study is the interlingual mapping of lexical meaning;
- have a *public and well-defined model* of lexical meaning that makes it possible to perform a formal analysis of lexical expressivity;
- *target natural languages*, as cross-lingual practices around specialized (domain) terminology and encyclopedic knowledge are different from general language and are out of scope for this work.

Thus, we do not consider in our study otherwise remarkable resources such as Wiktionary[3] (as it is lacking a formal representation of lexical meaning, a model for meaning-based interlingual mapping and, more generally, a formal structure), Glosbe[4] or PanLex[5] (as their internal representation of meaning is not fully public), DBpedia[6] or ConceptNet[7] as they are encyclopedic rather than lexical databases. Nor do we consider terminologies such as Agrovoc[8] as phenomena of linguistic diversity within specialised vocabularies is not the topic of our research.

We review and compare *EuroWordNet*, *BalkaNet*, the *Multilingual Central Repository*, two versions of the *Open Multilingual Wordnet*, *IndoWordNet*, *BabelNet*, and the *Universal*

---

[3] http://www.wiktionary.org.

[4] http://www.glosbe.com.

[5] http://www.panlex.org.

[6] http://www.dbpedia.org.

[7] https://conceptnet.io.

[8] https://agrovoc.fao.org.

Table 2 Comparison of the support of interlingual meaning representation and mapping features among MLDBs, as defined in Sect. 2

MLDB	Unbiased	Equivalence	Hypernymy	Untranslatability
EuroWordNet	No	Partial	Partial	Partial
BalkaNet	No	Partial	Partial	Partial
MCR	No	Partial	Partial	No
OMW	No	Partial	Partial	No
OMW2	Yes	Yes	Partial	Yes
IndoWordNet	Yes	Partial	Partial	No
BabelNet	Yes	Yes	Yes	No
UKC	Yes	Yes	Yes	Yes

Knowledge Core, showing how they take markedly different approaches to modelling cross-lingual mappings. Each review consists of a structural overview and an analysis of mapping ability based on a complex example of interlingual mappings around cousin-like family relationships. Table 2 provides a summary comparison according to the four criteria defined in Sect. 2.

All MLDBs studied formally distinguish between *words* and *word meanings*, as the correspondence between the two is often one-to-many (polysemy) or many-to-one (synonymy). For a coherent representation of different MLDBs, in the rest of the paper we adopt the *WordNet* model of word meanings and the corresponding terminology, introduced by Miller (1998); Fellbaum and Vossen (2007) and today used in thousands of wordnets and similar resources. In wordnets, lexemes are called *words* (even for multiword expressions). A word with a specific meaning is called a *sense*. The senses of synonymous words are linked to a single *synset* (synonym set) that formally represents the synonymous senses as collapsed into a single node. Synsets are interconnected into a graph through hierarchical relations of (intra-lingual) hypernymy and hyponymy (broader and narrower meaning), as in traditional thesauri.

3.1 EuroWordNet, BalkaNet, MCR, Open Multilingual Wordnet v1 & v2

Due to the many shared features, this section describes together EuroWordNet (EWN) (Díez et al. 1997; Vossen 1998), BalkaNet (Tufis et al. 2004), the Multilingual Central Repository (MCR) (Aitor Gonzalez-Agirre and Rigau 2012), as well as two versions of the *Open Multilingual Wordnet* (OMW and OMW2) (Bond and Paik 2012; Bond and Foster 2013; Bond et al. 2020). The EuroWordNet project pioneered the creation of multilingual wordnet resources and their cross-lingual mappings. It directly or indirectly influenced other collaborative efforts, under the umbrella of the *Global WordNet Association* (Vossen et al. 2016; Pease et al. 2008),[9] on specific language groups such as BalkaNet for the Balkans and MCR for the languages of Spain. The OMW, in turn, harmonised the representations of these and many other wordnets, e.g. Black et al. (2006) and Balkova et al. (2004), mapped all of them to the English Princeton WordNet 3.0 and, in its *Extended* version, expanded linguistic coverage to hundreds of languages with words automatically extracted from Wiktionary and the *Unicode Common Locale Data Repository*.

---

[9] http://globalwordnet.org/.
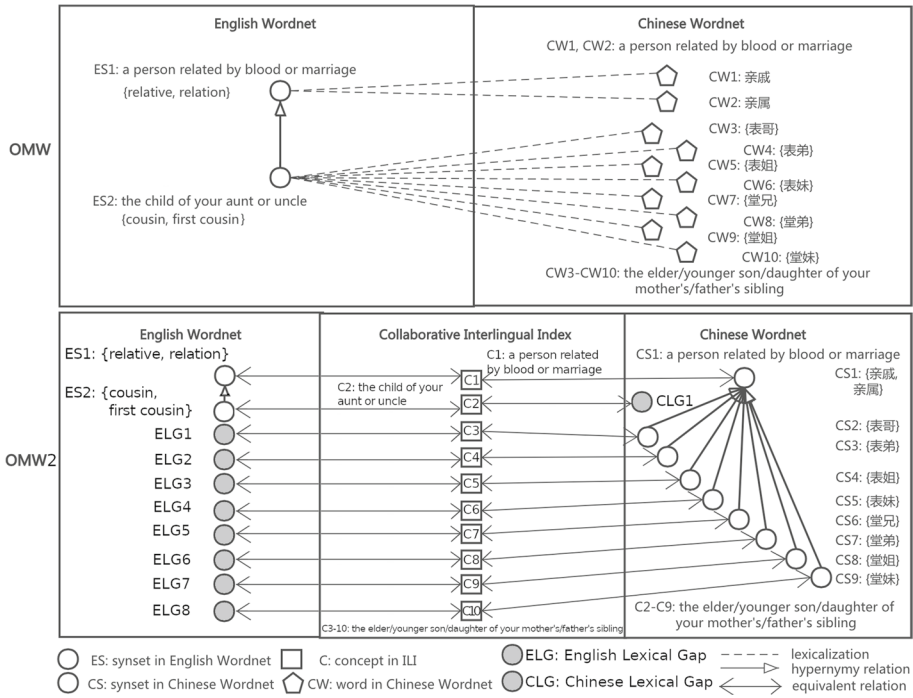
Fig. 1 Examples of Chinese-to-English mappings using the EuroWordNet/BalkaNet/OMW data models (top) and OMW2 (bottom)

All of these efforts use the English Princeton Wordnet (PWN) as their inter-lingual hub. EuroWordNet and BalkaNet link the synsets of separate language-specific wordnets to English PWN synsets through equivalence relations. MCR and OMW, on the other hand, link English synsets directly to words in other languages through *lexicalization* relations that, in practice, still imply meaning equivalence. In both cases, the use of PWN as a hub results in a bias towards the English language and culture: our criterion 1 on an unbiased meaning space is not fulfilled. Accordingly, MCR and OMW do not contain any word that has no equivalent English meaning in PWN. Some wordnets from EWN (e.g. Dutch) and BalkaNet (e.g. Romanian, Czech) contain language-specific synsets and lexical gaps, but the synsets are not mapped to other languages and the gaps are only mapped to English (hence the "partial" support for untranslatability in Table 2).

Figure 1 shows an example of Chinese-to-English mapping in OMW (the EWN/MCR/BalkaNet models behave the same way). The Chinese word CW1 is correctly mapped to the English meaning ES1 {*relative, relation*}. The eight Chinese words representing cousins are, however, all mapped to the single PWN synset meaning *cousin*. This results in a representation that is both incomplete and incorrect: the meanings of the more specific Chinese words are lost, while the mappings give the impression that these words are all synonyms and equivalent in meaning to the English *cousin*. The fact that these resources cannot express that the Chinese terms are more specific than *cousin* means that our criterion 3 on hypernymy is only partially fulfilled. Likewise, neither equivalence not untranslatability can be expressed for meanings not present in English (such as the ones in Table 1).

More recently, efforts towards a second version of OMW were announced (Bond et al. 2020). Even though, to our knowledge, as of early 2023, no dedicated lexical content distinct from that of OMW1 has been released for OMW2, we review the abilities of this database based on information available from the publications cited. OMW2 replaces the lexicalisation mappings of OMW (that relate English PWN synsets with lexicalisations from other languages) by synset-to-synset mapping relations towards a *Collaborative Interlingual Index* (CILI). The CILI is a set (i.e. an unstructured collection) of unique IDs that represent word meanings relevant to one or more languages. IDs within the CILI are linked to synsets within wordnets with one-to-one equivalence relations (implemented as `owl:sameAs` in the Semantic Web representation of the OMW2). The collaboratively-built and managed CILI is meant to expand beyond PWN to cover synsets that have no English equivalents, and thus eliminate the English-centeredness of OMW. OMW2 also introduces lexical gaps in order to distinguish between resource incompleteness and untranslatability.

Figure 1 shows the same *cousin* example as it can be modeled by OMW2. It allows the creation of new IDs within the CILI for the eight specific kinds of Chinese cousins, which can then be linked to other languages, or represented as lexical gaps. The eight Chinese meanings can thus be included in the CILI and their absence from the English vocabulary can be explicitly marked. Criteria 1, 2, and 4 (on the unbiased meaning space, equivalence, and untranslatability) are thus fulfilled. Note, however, that the graph in Fig. 1, composed of hypernymy edges within the wordnets as well as of equivalence relations towards the CILI, does not provide any relationship between the English meaning of *cousin* (ES2) and the more specific Chinese words (CS2–CS9). The fact that *cousin* is more general than CS2–CS9 is an example of *interlingual knowledge* that is not directly derivable from the union of monolingual lexicons and the CILI. Even if one wanted to represent this knowledge, it would not be possible within the OMW2 model using the CILI and equivalence mappings alone. As the CILI layer leaves hierarchical structuring of word meanings to individual wordnets, it cannot express cross-lingual hierarchical relationships. Criterion 3 on interlingual hypernymy is therefore only partially fulfilled.

3.2 IndoWordNet

IndoWordNet[10] (IWN) includes 18 languages from the Indo-Aryan, Dravidian, and Sino-Tibetan families (Dash et al. 2017; Bhattacharyya 2010; Singh et al. 2016; Kanojia et al. 2018; Saraswati et al. 2010). Similarly to other wordnets, IWN uses synsets to represent word meanings along with their associated glosses. One of the particularities of IWN is its use of the *Hindi WordNet* (HWN) (Narayan et al. 2002; Chakrabarti and Bhattacharyya 2004), as opposed to English, as the central hub that interconnects the 18 languages. Within IWN, only the HWN contains a synset hierarchy: the other 17 languages are represented as flat lists of synsets. The use of HWN (as opposed to PWN) as the hub makes sense for reasons of cultural and linguistic proximity to other languages of India. Accordingly, the HWN contains many synsets culturally and linguistically relevant to the Indian subcontinent.

While the limitation of word meanings to what is lexicalized in Hindi restricts the expressivity of IWN, the database does allow the creation of synsets specific to each of its 17 languages covered. Thus, IWN fulfils our criterion 1 on having an unbiased meaning

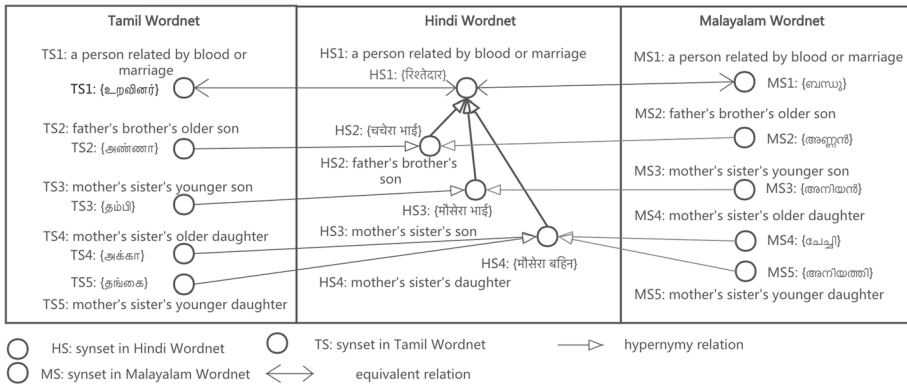---

[10] https://tdil-dc.in/indowordnet/.

Fig. 2 Example of Tamil–Hindi–Malayalam mappings using the IndoWordNet model

space. However, such language-specific meanings are not part of the hub which is limited to Hindi. Interlingual equivalence mappings therefore are limited to what is expressed by the Hindi lexicon.

This limitation is counterbalanced by the ability of IWN—unique among the resources reviewed—to use both equivalence and hypermymy for interlingual mapping. Figure 2 shows our *cousin* mappings between Hindi and Malayalam, a Dravidian language from Southern India. In Malayalam, MS1 can be mapped to HS1 using equivalent mapping, but MS2–MS17 are more specific meanings than HS2–HS9 which do not exist in HWN. The solution of IWN is to link them to a more general synset with hypernymy relations: it maps HS2 (father's sister's son) in Hindi to two more specific Malayalam meanings, MS2 and MS3 (father's sister's elder/younger son) through two hypernymy relations. IWN is thus capable of correctly mapping non-equivalent synsets across languages. On the other hand, due to Hindi being the hub, IWN is not able to map equivalent meanings across Indian languages if the meaning is not part of Hindi. For example, Tamil and Malayalam have lexicalizations for *mother's sister's elder daughter* (TS4 and MS4, resp.), but the IWN can only indicate that they are both hyponyms of HS4, resulting in information loss. IWN thus only partially fulfils criteria 2 and 3 on unbiased equivalence and hypernymy mappings. Finally, the lack of modelling lexical gaps means that IWN fails our criterion 4 on untranslatability.

3.3 BabelNet

BabelNet[11] stands between a semantic network and a lexical database, covering terms of both lexicographic and encyclopaedic origin (Navigli and Ponzetto 2012; Ehrmann et al. 2014). Version 5.2 of BabetNet contains 520 languages, and 22 million entries. Its contents were imported from online encyclopaedias and lexical resources such as wordnets, Wiktionary, Wikipedia, OmegaWiki, and Wikidata, which explains its larger size and wide coverage of named entities.

---

[11] https://babelnet.org/.
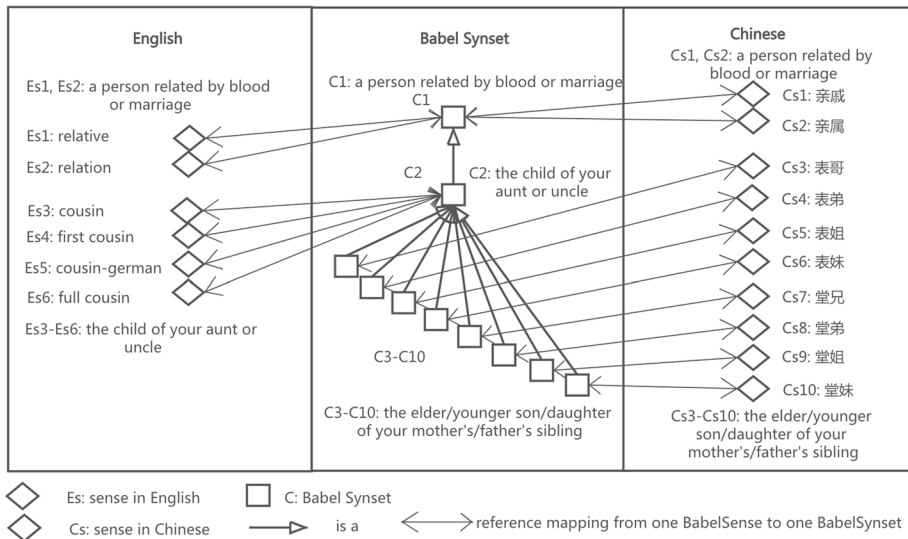
Fig. 3 Chinese-to-English mappings using the BabelNet model

BabelNet builds a unified, supra-lingual lexical meaning space, represented as a hierarchy of *BabelSynsets*. These, in turn, are lexicalized in each language by language-specific *BabelSenses*. As the synset hierarchy is defined outside of the language-specific lexicons, it becomes theoretically possible to build a meaning space unbiased towards any particular language. Figure 3 shows how our running example of English–Chinese mappings could in theory be represented in BabelNet. The supra-lingual central layer is capable of representing shared meanings (e.g. C1) as well as language-specific meanings (C2–C10), within a single hierarchy. Individual BabelSynsets are then mapped to one or more synonymous lexicalisations (BabelSenses) in each language. The model of BabelNet thus allows word meanings to be hierarchically related across languages (such as the English *cousin* and the eight more specific Chinese meanings), which is not possible for the DBs described in Sect. 3.1. It also avoids the limitation of IWN of not being able to map meanings that are not in the hub language. BabelNet thus fulfils criteria 1 to 3, but not criterion 4 as it does not offer any information on untranslatability.

In practice, however, BabelNet does not exploit its structural potential to address language diversity explicitly. This becomes clear by observing how BabelNet actually represents the eight Chinese meanings CS3–CS10: in contrast to the correct mappings shown in Fig. 3, of which BabelNet is theoretically capable, it maps most of them to the PWN meaning of *cousin* and leaves the remaining ones unmapped.

3.4 The universal knowledge core

The universal knowledge core (UKC) (Giunchiglia et al. 2017, 2018) is a large-scale MLDB that contains about 2 million words in over 2000 languages (Bella et al. 2022b).[12] It integrates a variety of resources such as individual wordnets such as (Ganbold et al. 2018;

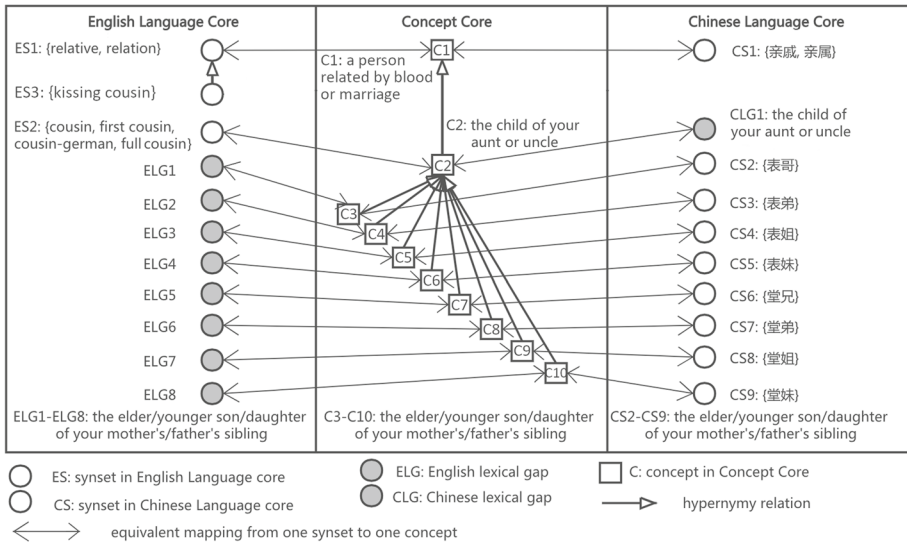[12] http://ukc.datascientia.eu/.

Fig. 4 Example of English–Chinese mappings as supported by the UKC

Bella et al. 2020), Wiktionary, as well as original multilingual content on phenomena related to linguistic diversity, such as cognacy (Batsuren et al. 2022), metonymy (Khishigsuren et al. 2022b), lexical gaps (Khishigsuren et al. 2022a), morphology (Batsuren et al. 2021), lexical similarity (Bella et al. 2021). The UKC has a two-layered architecture, with a *language layer* that contains a separate wordnet-like graph (with words, senses, and synsets) for each language, as well as a supra-lingual layer of *interlingual concepts*Giunchiglia et al. (2018) (Fig. 4). Each such concept represents a word meaning from at least two of the constituting languages, so that the concept layer consists of the *union* of all word meanings that are mapped to at least one other language. Thus, in our running example, each of the eight Chinese meanings of *cousin*, the eight Hindi meanings, and the 16 Malayalam meanings becomes a separate interlingual concept. The UKC thus has an unbiased meaning space (criterion 1).

Yet, the UKC does not assume that lexical meaning within all languages can be perfectly described with a single unified concept graph. A major distinguishing feature with respect to all previously presented MLDBs is the ability to represent word meanings and their hierarchy *both* on the interlingual and on the language-specific levels, the former using concepts and the latter synsets. Thus, we allow smaller unaligned hierarchies to coexist with the merged core of interlingual meanings. This architectural choice reacts to the impossibility of ever reaching a perfect merge of *all* lexicons for *all* languages of the world, both due to the effort implied and allowing for irreducible cases of diversity. For example, in Fig. 4, the newly introduced culture-specific English *kissing cousin*, meaning *a relative with whom someone is in kissing terms*, may need to be aligned with concepts from other languages before it can be integrated into the concept layer, and is thus temporarily kept as a synset-level meaning within the English language layer, all the while being linked to concepts in the overall UKC graph through hypernymy.

Interlingual equivalence is represented in the UKC by mapping language-specific synsets to the same concept. For example, the UKC maps the English synset {*relative, relation*}, the Italian {*parente, familiare*}, and the Chinese {亲戚,亲属} to the same

interlingual concept. Thus, the interlingual concept layer acts as the hub and the UKC, just like OMW2, is capable of representing equivalence mappings (criterion 2).

Interlingual hypernymy and hyponymy are represented within the concept layer. In this respect, the UKC is different from OMW2 which keeps meaning hierarchies within the original resources. Representing all word meanings as well as their relationships in a single graph means that, as in the case of BabelNet, any pair of word meanings can be put in a broader–narrower relation (criterion 3).

Untranslatability, finally, has explicit support in the UKC through the *lexical gap synset* that, contrary to regular synsets, does not have senses or words attached to it, but does have a gloss. When a concept is not lexicalized in a language, it is mapped to a lexical gap synset instead of leaving it unmapped (as shown in Fig. 4). This feature allows for distinguishing resource incompleteness from untranslatability (criterion 4).

The ability of the UKC to represent interlingual equivalence, hypernymy, and untranslatability can be exploited in computational applications such as machine translation or cross-lingual transfer learning, in order to improve their precision in linguistically diverse domains. For example, when translating the Chinese 堂妹 (younger female patrilineal cousin) to English, a machine translation system can be informed by the UKC that the Chinese word has no English equivalent (it is a gap in English), but that a broader English word *cousin* exists, which is the most suitable single-word translation available. This operation is not symmetric: 堂妹 should not be automatically considered as a correct translation for *cousin*, as it implies additional information that may be wrong depending on the context.

4 A quantitative evaluation

We evaluate and compare the MLDBs presented in Sect. 3 in terms of our four criteria on interlingual mapping ability: how the structure of each resource determines its *coverage* of language-specific concepts, interlingual equivalence, hyper/hyponymy, and untranslatability mappings.

4.1 Evaluation data

As the focus of this paper are the structural abilities of MLDBs rather than the completeness of their actual content—which varies to a great degree according to the languages covered—we evaluate mapping expressivity on an ad-hoc gold standard set of interlingual mappings. The dataset consists of $|C| = 288$ lexical concepts (language-specific word meanings) that include 160 lexicalizations and 128 lexical gaps from nine languages and five phyla (English, French, Italian, Chinese, Hindi, Tamil, Malayalam, Hungarian, and Mongolian), all provided by native speakers. The words were deliberately selected from five culturally diverse semantic groups, belonging to four distinct domains: words expressing various *kinship* relations (siblings, cousins, elder/younger, male/female, etc.), kinds of *watercourses* (according to size), *horses* (male/female, young/adult), and *rice* (raw/cooked, white/brown, cleaned or in the husk). The gold standard set contained the exhaustive mappings within each semantic group, in terms of equivalences, $R_{\equiv}(C) = 431$, hyper/hyponymy, $R_{\sqsubset}(C) = 1139$, and untranslatability, $R_{\text{GAP}}(C) = 389$, totalling in 1959 gold-standard interlingual mapping relations. The Online Appendix provides the complete list of words and gaps, as well as details on corpus development.

Table 3 Interlingual concept and mapping coverage for each MLDB evaluated

Criterion	Gold std.	OMW-like (%)	OMW2 (%)	IWN (%)	BbN (%)	UKC (%)
1. Concept coverage	32	56.3	100	100	100	100
2. Equivalence rel	431	88.9	100	77.3	100	100
3. Hyper+hypo rel	1139	61.6	80.9	90.1	100	100
4. Untranslatability rel	389	0	100	0	0	100
Total relations	1959	55.4	88.9	69.4	80.1	100

4.2 Evaluation method

The evaluation consisted of manually analyzing the representational ability of MLDBs against each mapping. We included OMW2, IWN, BabelNet, the UKC, and the OMW, the last one equivalent in its mapping abilities to EWN, MCR, and BalkaNet and thus representative of them as well. This involved the analysis of $1,959 \times 5 = 9795$ mapping instances.[13] The Online Appendix gives more detail on how the evaluation of MLDBs was performed against the gold standard corpus.

In order to compute coverage results in Table 3, we defined the *interlingual concept coverage* $CCvg(C, \mathcal{D})$ of an MLDB $\mathcal{D}$ with respect to a set of lexical concepts C in the following very simple way:

$$CCvg(C, \mathcal{D}) = \frac{|C^{\mathcal{D}}|}{|C|},$$

where $C^{\mathcal{D}} \subseteq C$ are the concepts from C that $\mathcal{D}$ is able to express. In a similar manner, we defined the *interlingual mapping coverage* $MCvg(r, C, \mathcal{D})$ of an MLDB $\mathcal{D}$ with respect to the same set of lexical concepts C and the mapping relation type r as follows:

$$MCvg(r, C, \mathcal{D}) = \frac{|R_r^{\mathcal{D}}(C)|}{|R_r(C)|} \text{ where } R_r^{\mathcal{D}}(C) \subseteq R_r(C) \subseteq C \times C,$$

where $r \in \{\equiv, \sqsubseteq, \sqsupseteq, GAP\}$, i.e. one of the mapping relationships evaluated throughout our paper, $R_r(C)$ is the set of all correct interlingual relations of type r over the set of concepts C, and $R_r^{\mathcal{D}}(C)$ is a subset of these relations that $\mathcal{D}$ is able to express.

Quantitative results can be found in Table 3. In the following we provide both a discussion of the results.

4.3 Discussion

All MLDBs evaluated, except for OMW1 (and the similar EWN, BalkaNet, and MCR), provide a mechanism for adding language-specific concepts to the database. OMW1, instead, is limited to the synsets present in the English WordNet, which covers only 18 concepts out of 32 in our gold standard, corresponding to the concept coverage of 56.25% shown in Table 3.

---

[13] As we were interested in the structural properties of IWN, we made abstraction of its limitation to Indian languages.

All MLDBs generally support equivalence mappings and were able to express most of such mappings in our test set. OMW-like databases and IWN, however, are unable to express equivalences that involve meanings that are missing from their hub language (English and Hindi, resp.), such as $fleuve_{\text{FRENCH}} \equiv folyam_{\text{HUNGARIAN}}$ (meaning a particularly large river) or $mchele_{\text{SWAHILI}} \equiv$ 生米$_{\text{CHINESE}}$ (meaning uncooked rice). This is a form of structural bias. OMW2, BabelNet, and the UKC, on the other hand, are able to represent all equivalences through their extensible hubs that create a node (a CILI entry, a BabelSynset, and a concept, respectively) for each word meaning lexicalized in at least one language.

In terms of interlingual hypernymy mappings, larger differences are observed among the MLDBs. While BabelNet and the UKC are able to express 100% of our test set mappings, the remaining resources are weaker. In the case of OMW, EWN, BalkaNet, and MCR, only the PWN-based hub contains a hierarchy, which means that these resources can only express such relations if they are also present in the PWN. Thus, these MLDBs miss 38.4% of hypernymy and hyponymy from our test set. OMW2 takes the opposite approach and relies on the individual wordnet hierarchies and the cross-lingual equivalence mappings (as shown in Fig. 1) to infer them. This is not sufficient to compute certain mappings, such as the relation between the English *cousin* and the more specific Malayalam words, as the meaning of cousin is a lexical gap in Malayalam. IWN, in turn, is more powerful due to its use of cross-lingual hypernymy mapping relations, and is therefore able to express the English–Malayalam relation (as well as many others) via hypernymy through a Hindi hub meaning. Yet, it would not be able to express hypernymy between *cousin* and the Chinese 表姐 as no relation exists between the Chinese and any of the Hindi meanings. BabelNet and the UKC were able to express all mappings as they foresee the creation of a hub concept for each meaning and, contrary to OMW2, define the hierarchy within their hubs.

Finally, for untranslatability mappings, only OMW2 and the UKC provide explicit support for lexical gaps; this is visible from the table within Fig. 3. All other resources confound gaps with incompleteness, not differentiating a gap from a missing mapping.

4.4 Study limitations

As stated earlier, our goal was to quantitatively evaluate the impact of the theoretical mapping abilities of MLDBs on their coverage of a gold-standard interlingual mapping space. The abilities of each MLDB were formalised in our evaluation based on an analysis of their contents (when available) as well as on their descriptions in publications. While we do provide general qualitative information on the actual contents of each MLDB, these contents were not used in our evaluations.

Our evaluation covered concepts taken from four domains well known for their cross-lingual diversity: kinship, animals, geography, and food. We do not expect that the inclusion of new diversity-rich domains, such as colors or body parts, would affect our analysis and qualitative findings. That said, a less varied choice of domains and languages (e.g. the inclusion of more European languages or of lexically more uniform domains such as mathematics) would certainly lead to more homogeneous results in mapping abilities. Our evaluation languages and domains was admittedly and deliberately selected in order to amplify phenomena of lexical diversity as much as possible.

5 Conclusion

In this paper we dealt with the problem of how language diversity is represented in state-of-the-art multilingual lexical databases, an important issue in a globalized world where multilingual interactions are the norm and where, at the same time, the vast majority of languages does not benefit from adequate digital support. Current MLDBs should, at the minimum, leave open the possibility for these languages to integrate with the others, all the while avoiding any loss in their capacity of expressing lexical meaning specific to them. Our analysis, consisting of a theoretical qualitative and an example-based quantitative part, has shown largely differing cross-lingual mapping abilities among the MLDBs examined. We were able to explain these findings by the various ways of MLDBs to define language-specific meaning and their differing support of interlingual mapping.

References

Adamska-Sałaciak A (2010) Examining equivalence. Int J Lexicogr 23(4):387–409

Aitor Gonzalez-Agirre EL, Rigau G (2012) Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In: Proceedings of the 6th Global WordNet conference

Balkova V, Sukhonogov A, Yablonsky S (2004) Russian wordnet. In: Proceedings of the Second Global Wordnet conference

Batsuren K, Bella G, Giunchiglia F (2021) Morphynet: a large multilingual database of derivational and inflectional morphology. In: Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology. pp 39–48

Batsuren K, Bella G, Giunchiglia F (2022) A large and evolving cognate database. Lang Resour Eval 56(1):165–189

Bella G, McNeill F, Gorman R et al (2020) A major wordnet for a minority language: Scottish gaelic. In: Proceedings of the 12th language resources and evaluation conference. pp 2812–2818

Bella G, Batsuren K, Giunchiglia F (2021) A database and visualization of the similarity of contemporary lexicons. In: International conference on text, speech, and dialogue. Springer, pp 95–104

Bella G, Batsuren K, Khishigsuren T et al (2022a) Linguistic diversity and bias in online dictionaries. University of Bayreuth African Studies Online. p 173

Bella G, Byambadorj E, Chandrashekar Y et al (2022b) Language diversity: Visible to humans, exploitable by machines. In: Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations. pp 156–165

Bentivogli L, Pianta E (2000) Looking for lexical gaps. In: Proceedings of the ninth EURALEX international congress. Universität Stuttgart, Stuttgart, pp 8–12

Bhattacharyya P (2010) Indowordnet. In: In Proceeding of LREC-10, Citeseer

Black W, Elkateb S, Rodriguez H, et al (2006) Introducing the Arabic wordnet project. In: Proceedings of the third international WordNet conference, Citeseer. pp 295–300

Bond F, Foster R (2013) Linking and extending an open multilingual wordnet. In: Proceedings of the 51st annual meeting of the association for computational linguistics, vol. 1. pp 1352–1362

Bond F, Paik K (2012) A survey of wordnets and their licenses. Small 8(4):5

Bond F, da Costa LM, Goodman MW et al (2020) Some issues with building a multilingual wordnet. In: Proceedings of The 12th language resources and evaluation conference. pp 3189–3197

Catford JC (1978) A linguistic theory of translation. Oxford University Press, Oxford

Chakrabarti D, 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

Dash NS, Bhattacharyya P, Pawar JD (2017) The WordNet in Indian languages. Springer, New York

Díez P, Peter W, Vossen P (1997) The multilingual design of eurowordnet. In: Proceedings of ACL/EACL-97. Workshop on automatic information extraction and building of lexical semantic resources for NLP applications. Madrid

Eberhard DM, Simons GF, Fennig CD (2022) Ethnologue: languages of the world, 25th edn. SIL International. https://www.ethnologue.com/

Ehrmann M, Cecconi F, Vannella D et al (2014) Representing multilingual data as linked data: the case of babelnet 2.0. In: Chair NCC, Choukri K, Declerck T et al (eds) Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik

Fellbaum C, Vossen P (2007) Connecting the universal to the specific: towards the global grid. In: International Workshop on intercultural collaboration. Springer, pp 1–16

Ganbold A, Chagnaa A, Bella G (2018) Using crowd agreement for wordnet localization. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC-2018)

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

Giunchiglia F, Batsuren K, Freihat AA (2018) One world–seven thousand languages. In: Proceedings 19th international conference on computational linguistics and intelligent text processing, CiCling2018, 18–24 March 2018

Gurevych I, Eckle-Kohler J, Matuschek M (2016) Linked lexical knowledge bases: foundations and applications. Synth Lect Human Lang Technol 9(3):1–146

Joseph H, Heine SJ, Ara N (2010) The weirdest people in the world? Behav Brain Sci 33(2–3):61–83

Kanojia D, Patel K, Bhattacharyya P (2018) Indian language wordnets and their linkages with Princeton wordnet. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)

Khishigsuren T, Bella G, Batsuren K et al (2022a) Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. Preprint at http://arxiv.org/abs/2204.05049

Khishigsuren T, Bella G, Brochhagen T et al (2022b) Metonymy as a universal cognitive phenomenon: evidence from multilingual lexicons. In: Proceedings of the 44th annual conference of the Cognitive Science Society

Kornai A (2013) Digital language death. PLoS ONE 8(10):e77056

Lehrer A (1970) Notes on lexical gaps. J Linguist 6(2):257–261

Miller GA (1998) WordNet: an electronic lexical database. MIT Press, Berlin

Narayan D, Chakrabarti D, Pande P et al (2002) An experience in building the indo wordnet—a wordnet for Hindi. In: First international conference on global wordnet, Mysore, India

Navigli R, Ponzetto SP (2012) Babelnet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 193:217–250

Oxford Internet Study (2015) The digital language divide. http://labs.theguardian.com/digital-language-divide/

Pease A, Fellbaum C, Vossen P (2008) Building the global wordnet grid. CIL18

Saraswati J, Shukla R, Goyal RP et al (2010) Hindi to english wordnet linkage: challenges and solutions. In: Proceedings of 3rd IndoWordNet workshop, international conference on natural language processing 2010 (ICON 2010)

Singh M, Shukla R, Saraswati J et al (2016) Mapping it differently: a solution to the linking challenges. In: Eighth global wordnet conference

ten Hacken P (2016) Bilingual dictionaries and theories of word meaning. In: Proceedings of the XVII EURALEX International Congress, Lexicographic Centre, Ivane Javakhishvili Tbilisi State University Tbilisi. pp 61–76

Tufiş D, Cristea D, Stamou S (2004) Balkanet: aims, methods, results and perspectives. A general overview. Rom J Inf Sci Technol 7(1–2):9–43

Vossen P (1998) Introduction to eurowordnet. In: EuroWordNet: a multilingual database with lexical semantic networks. Springer, p 1–17

Vossen P, Bond F, McCrae J (2016) Toward a truly multilingual global wordnet grid. In: Proceedings of the eighth global WordNet conference. pp 25–29