Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2798–2807
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship

Temuulen Khishigsuren[1], Gábor Bella[2], Khuyagbaatar Batsuren[1], Abed Alhakim Freihat[2], Nandu Chandran Nair[2], Amarsanaa Ganbold[1], Hadi Khalilia[2], Yamini Chandrashekar[2], Fausto Giunchiglia[2]
[1]National University of Mongolia, Mongolia, [2]University of Trento, Italy
kh.temulen@gmail.com, gabor.bella@unitn.it, khuyagbaatar@num.edu.mn,
abdel.fraihat@gmail.com, nandu.chandrannair@unitn.it, amarsanaag@num.edu.mn,
hadi.khalilia@unitn.it, yamini.chandrashekar@unitn.it, fausto.giunchiglia@unitn.it

Abstract

This paper describes a method to enrich lexical resources with content relating to linguistic diversity, based on knowledge from the field of lexical typology. We capture the phenomenon of diversity through the notions of *lexical gap* and *language-specific word* and use a systematic method to infer gaps semi-automatically on a large scale. As a first result obtained for the domain of kinship terminology, known to be very diverse throughout the world, we publish a lexico-semantic resource consisting of 198 domain concepts, 1,911 words, and 37,370 gaps covering 699 languages. We see potential in the use of resources such as ours for the improvement of a variety of cross-lingual NLP tasks, which we demonstrate through a downstream application for the evaluation of machine translation systems.

Keywords: lexical typology, multilingual resource, lexical gap, kinship, linguistic diversity, multilingual NLP

1 Introduction

To address the language technology bottleneck beyond a handful of well-supported languages, the computational linguistics community has proposed several solutions: unsupervised learning models that remove the need for parallel textual corpora (Snyder and Barzilay, 2008; Artetxe et al., 2017; Radford et al., 2019) or cross-lingual transfer from high- to low-resourced languages (Hwa et al., 2005; Padó and Lapata, 2005; Täckström et al., 2012). Joint supervised learning and representation learning also happen to be effective in certain multilingual NLP applications, e.g. neural machine translation (Ammar et al., 2016; Guo et al., 2016).

The common point of these methods is that they rely on an implicit assumption of *sameness* across languages: for example, that in cross-lingual transfer the lexicon of one language can be efficiently mapped to that of another language through a shared semantic vector space. While the existence of shared cross-lingual conceptualizations is obvious to all—otherwise no interlingual communication could ever be possible—the world's languages are, however, also known to be extremely *diverse* on every level (Levinson and Evans, 2010). Thus, in recent years there has been an emerging trend to exploit results from the field of *linguistic typology* into language diversity in various multilingual NLP tasks (for a review, please see Ponti et al. (2019) and Arora et al. (2022)). Typology-informed NLP studies use typological features from hand-curated resources, such as the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), to bind parameters of languages that have similar typological features within the framework of cross-lingual transfer (Täckström et al., 2012) or to manipulate joint supervised learning models to reflect typological features of specific languages (Ammar et al., 2016; Bjerva et al., 2020). The use of cross-lingual cognate databases for bilingual lexicon induction was explored by Batsuren et al. (2021). The use of typology in various NLP applications generated consistent improvement, and is thus regarded as an effective way to advance the development of multilingual NLP (Cotterell et al., 2019; Salesky et al., 2020; Pimentel et al., 2021).

Most typology-informed NLP studies, however, are limited to recognizing language-specific morphosyntactic features and have so far ignored diversity within lexicons. Yet, the vocabularies of languages differ considerably in their division of the semantic space. For example, many languages lack an equivalent to the English word *cousin*, and instead employ several (up to 16) more specific words that distinguish male and female cousins, elder and younger, paternal and maternal, etc. The ignorance of such semantic diversity in language processing can lead to hard-to-detect meaning-level mistakes. Today's top machine translation systems, for instance, make consistent mistakes over simple sentences such as *My brother is younger than me* even across relatively high-resource languages: for example, the Japanese translation of the English sentence above is 私の兄は私より若いです with a nonsensical meaning of *My elder brother is younger than me.* The same mistake is observed over many other languages such as Mongolian or Hungarian.

It is clear from the example above that manifestations of language diversity are only revealed explicitly in a cross-lingual context, and remain hidden as long as one's point of view remains monolingual. Hence, the way to formalise lexical diversity is through building *typology-informed multilingual lexical resources*. In this paper we exploit existing knowledge from lexical typology to enrich and extend existing multilingual lexicons which, in turn, can be reused in NLP applications. The key notion to capture lexical diversity is that of the *lexical gap*, which refers to the lack of lexicalization of a particular concept in a particular language. We provide a formal approach to infer lexi-

cal gaps semi-automatically from existing domain-specific lexical knowledge. We apply the method to the domain of *kinship*, well known for its cross-lingual diversity, building a resource that provides 37,370 gaps in 699 languages. Finally, as an example application in the context of multilingual NLP, we demonstrate how our resource can be used to evaluate machine translation systems over semantically challenging corpora. To our knowledge, ours is the first systematic method and the first lexical resource that provides lexical gaps for multiple languages in a large-scale and exhaustive manner.

The rest of the paper is organized as follows. We start by defining the notion of lexical gap and giving a brief review of lexical typology in Section 2. Our typology-based lexical gap generation method is described in Section 3. Section 4 evaluates generation results based on native speaker input. Sections 5 and 6 present the resulting dataset and an example application for improving machine translation. Finally, we review related work and provide conclusions in sections 7 and 8.

2 Untranslatability and Lexical Typology

The notion of *lexical gap* is closely related to that of *untranslatability* (Catford, 1978). The latter, however, is a practically-oriented concept with many possible definitions: the acceptability of a translation depends on how rigorous one is regarding precision, conciseness, register, style, etc. Accordingly, the notion of lexical gap can be defined in broader or narrower ways. Linguists tend to adopt a more strict definition where monomorphemic words for which an equally monomorphemic translation cannot be provided are considered as gaps. For example, Wierzbicka (2008) considered that the concept of "color" is a lexical gap in Warlpiri, an Australian Indigenous language, as it lacks a word for it. Another example from Joo (2021), the general concept of "rice" is a lexical gap in Korean, which instead has words for more specific concepts "cooked rice" and "uncooked rice". In a computational context, Bentivogli and Pianta (2000) adopt a somewhat more relaxed definition, distinguishing translations into single words and *restricted collocations* on the one hand, and *free combinations of words* on the other hand, considering only the former as valid lexicalizations. Under this account, a restricted collocation is a stable combination of words (e.g., "elder brother"). On the other hand, a free combination of words is a combination of words that are not bound together, and its parts can be freely used with other lexical items (e.g., "father's elder brother"). In our research, we adopted the definition of Bentivogli and Pianta (2000) and, accordingly, considered a concept to be a lexical gap in a language if it can only be expressed through a free combination of words. Another criterion was only to consider as lexicalizations general-language words understood by "average" native speakers, as opposed to specialized terminology or rare and unknown words. For example, the gender-independent English *nibling* and French *adelphe* are specialized terms that designate *nephew or niece* and *sibling*, respectively. As these neologisms are only understood by specialists and are never used by the general public, we considered them as lexical gaps.

Lexical typology, a field of linguistics, explores systematic variations in the presence or absence of lexicalizations in languages with respect to specific domains (also known as "domain categorization") (Koch, 2001). English and Northern Sami, for example, categorize uncle-like relationships differently: Northern Sami has three different words with three distinct meanings—*eahki* "father's elder brother," *čeahci* "father's younger brother," and *eanu* "mother's brother"—while English packs all these meanings into the single term *uncle*.

Studies in lexical typology have mostly been conducted on domains that offer an unexpected richness of cross-lingual diversity: body parts, color, kinship, perception verbs, motion events, spatial dimension terms, cardinal direction terms, cutting and breaking events, putting and taking events, or pain predicates (for a review, see Koptjevskaja-Tamm et al. (2015) and Arora et al. (2021)). However, only a small part of related scientific results have been published as actual datasets: Murdock (1970)'s kinship categorization is published in D-PLACE (Kirby et al., 2016). Parts of Brown (1976) and Kay et al. (2009)'s works on colors are published under the lexicon chapter of WALS. To the best of our knowledge, no other works have been published as open resources.

More recently, digital lexical resources and parallel corpora have been increasingly used in lexical typology, enabling typologists to cover more languages and domains. For example, Viberg (2014)'s 50-language study on the perception domain was extended to 1,220 languages in Georgakopoulos et al. (2021). A study of the color domain in 119 languages by Kay et al. (2009) was leveraged in McCarthy et al. (2019), which managed to cover 2,491 languages. Recent automated attempts to obtain lexico-semantic knowledge from large-scale parallel corpora provided promising results as well (Gast and Koptjevskaja-Tamm, 2018; Levshina, 2021). Such bottom-up, data-driven approaches extend existing knowledge on lexical typology which, in turn, broaden the applicability of our method.

3 Lexical Gap Generation Method

In this section, we describe a top-down, linguistically informed method for the systematic generation of lexical gaps for a large number of languages. While the method is generic and can be applied to multiple domains, our efforts so far have concentrated on the domain of kinship relations, from which all of our examples are taken. The method consists of the following three steps:

1. *domain specification:* the coverage of the study is defined in terms of domains and subdomains;

2. *conceptual modeling:* the domain selected is formalised through an interlingual hierarchy of lexical concepts;

3. *gap generation:* for each language covered by typological data, non-lexicalized concepts are marked as lexical gaps.

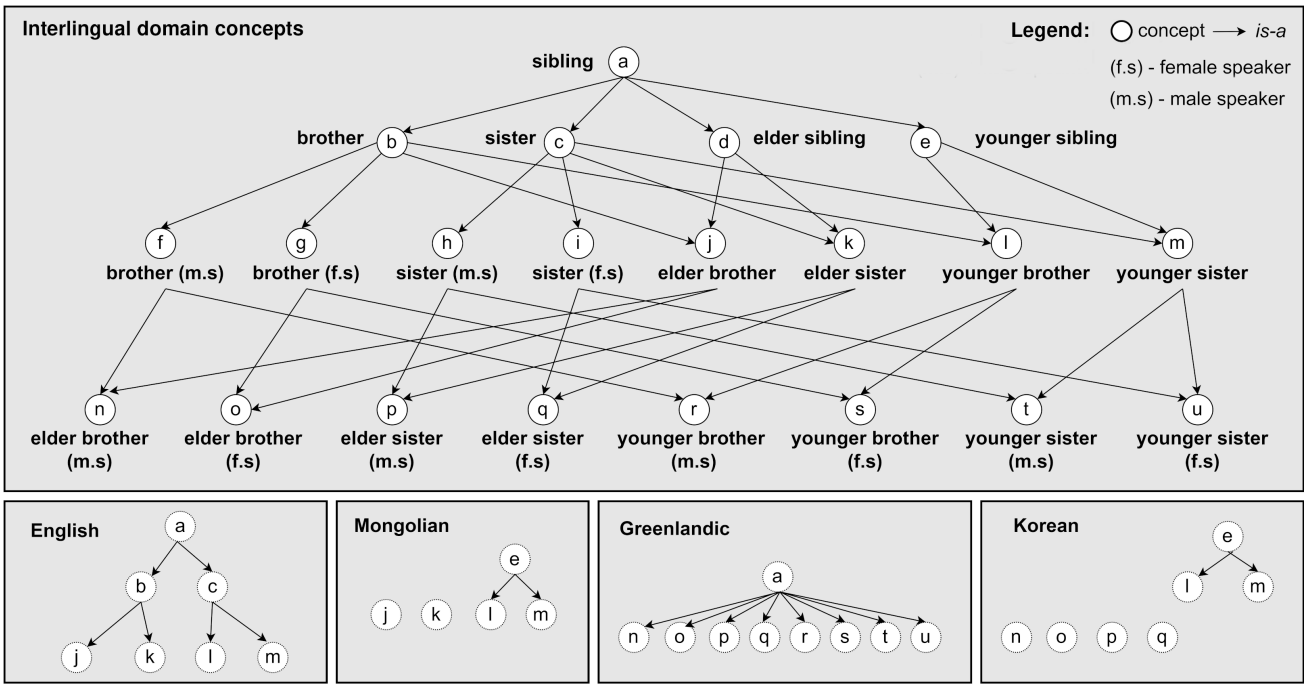

Figure 1: Interlingual conceptual layer of sibling domain.

3.1 Domain specification

This step defines the domain of interest, based on the availability of typological data (lexicalization patterns) for the languages to be covered. The larger the domain and its subdomains, the more complex it is in the subsequent step to provide an interlingual representation. We chose to represent the *kinship* domain as it is a crucial part of the general lexicon, yet it is known to be enormously diverse across languages and cultures. More precisely, we covered the six subdomains of *grandparents*, *grandchildren*, *siblings*, *uncles and aunts*, *nephews and nieces*, and *cousins*. We relied on the seminal work of Murdock (1970) for lexicalization patterns and for a general characterization of the domain.

3.2 Conceptual Modeling

Modeling the interlingual conceptual space is an essential part of inferring lexical gaps in a systematic way. A purely conceptual approach to modeling the interlingua—developing or adopting an actual domain ontology—while feasible with limited effort and linguistic knowledge, produces models that are overly complex with sparse mappings to languages (e.g. providing lots of non-lexicalized concepts). A purely linguistically motivated approach, instead, would produce results close to actual language use but would be extremely onerous, as it is based on the observation of potentially thousands of languages.

We propose a hybrid linguistic–conceptual approach where a top-down domain ontology is constrained by bottom-up linguistic data provided by the typological literature. In practice, for the kinship domain, we relied on Murdock (1970) as an authoritative source of linguistic domain knowledge. This work provided the key formal attributes to consider for building the interlingua: *gender of the kin* and *of the speaker* (male, female, undefined), *age of the kin* relative to the sibling and to the speaker (older, younger, undefined), *relationship* (parent, child, or sibling). In two subdomains out of six (nephew/niece and cousin), information from typology literature was incomplete: for these subdomains, we relied on complementary information sources, namely native speaker input and Wiktionary words.

From these attributes, we automatically generated an exhaustive set of concepts and *is-a* relations, such as *parent's male sibling* or *male parent's male sibling's female child (i.e. father's brother's daughter) as pronounced by a female speaker*. The first concept is lexicalised in English as *uncle* while the second one, a specific type of *cousin* relationship, in Aleut as *agiitudaxtanax*. There are, however, many theoretical combinations that are not actually lexicalized in any language. In order not to generate concepts that unnecessarily increase the size and complexity of the interlingua, we filtered the set of concepts generated to include only those attested by typological literature and native speaker input. As a result, we identified 198 concepts and 347 *is-a* relations covering the six subdomains. A simplified example of the coverage of the *sibling* subdomain can be seen in the top layer of Figure 1, with the bottom layer showing existing lexicalizations in four languages.

Source data	Languages	Words	Subdomains
Murdock (1970)	566	0	4
Wiktionary	166	1,681	6
Native speakers	10	230	6

Table 1: Types of data used to infer lexical gaps.

3.3 Gap Generation

For each concept defined above, this last step verifies whether it is lexicalized in each language covered, and if

not, a lexical gap in that language is generated. In the kinship domain, we used three sources of evidence for lexicalization: (1) Murdock's lexicalization patterns, (2) Wiktionary words, and (3) direct input from native speakers. Table 1 summarizes the types of data used to infer lexical gaps. (1) provided lexicalization patterns for 566 languages, which we retrieved from Kemp and Regier (2012). To give an example of a lexicalization pattern, the Javanese language belongs to the *Algonkian sibling pattern*, which means that it lexicalizes the sibling terms: *elder brother*, *elder sister*, and *younger sibling*. Although this database does not provide the actual words, it can be used to infer gaps. (2) From Wiktionary, we extracted kinship terms in 166 languages, 144 of which were complementary to the Murdock dataset, using a custom-written parsing script. From the terms extracted, we automatically inferred the presence of gaps using the inference rules presented below. (3) Lastly, for the languages Arabic, English, French, Hindi, Hungarian, Italian, Kannada, Malayalam, Mongolian, and Spanish, we asked two native speakers per language to provide lexicalization and gap information for all concepts. This information was used for evaluation (see Section 4 below) but also as a gold-standard source of terms and gaps for languages where incomplete or no data was available.

As Wiktionary does not explicitly indicate lexical gaps, we used the following rules to infer the presence of gaps from existing lexicalizations. The rules were derived from empirical patterns observed in the typology literature.

For concepts with speaker gender and age undefined: if neither a concept c nor its parents have a lexicalization in language l then c is a lexical gap in l.

For concepts with speaker gender or age specified: if language l is known not to indicate the speaker's gender or age in the lexicalization, then all concepts with these attribute are lexical gaps in l.

We explain the first rule through the example of *uncle*. If a language lexicalizes *uncle*, it might also express the more specific *paternal* and *maternal uncle*, e.g., by adding appropriate adjectives. We cannot automatically infer that *paternal uncle* and *maternal uncle* are gaps: deciding whether collocations are restricted is far from trivial, as even native speakers may disagree on the everyday usage of expressions such as *paternal uncle*, *younger sibling*, or *female cousin*.[1] We consider, however, that complex expressions corresponding to indirect hyponyms (of distance 2 or more), such as *father's elder brother* or *mother's younger sister*, are never restricted collocations and can therefore be considered as gaps.

The second rule is explained by the rarity of the phenomenon of incorporating the speaker's gender and age into lexicalizations (e.g. a male Ewe person would call his grandchild differently than a female Ewe would). It is safe to assume that languages that do not possess these properties have no concise way of expressing this information as part of the kinship terminology.

[1]The use of corpus-based frequency data is promising future work in this direction.

4 Evaluation

Our evaluations verified the correctness and completeness of the automated word extraction and gap generation method described in Section 3, and also extended the overall coverage of our resource. Our goal was to obtain highly reliable and reusable data on lexical diversity, we thus gave priority to precision over recall. Furthermore, as we considered our sources from linguistic typology as *a priori* reliable, we instead focussed our evaluations on the words and gaps that were automatically inferred from Wiktionary data. Thus, we verified the correctness of: (a) the original Wiktionary data, generally considered as reliable, nevertheless not fully error-free; (b) our Wiktionary data extraction logic; (c) our lexical gap inference rules.

Evaluation Setup

The evaluation was based on input provided by language speakers in ten languages: Arabic, English, French, Hindi, Hungarian, Italian, Kannada, Malayalam, Mongolian, and Spanish. Two speakers per language were employed: the first one a native speaker born and educated (university-level) within the speaker community, while the second speaker was every time a language expert with (at least) proficiency in the language and a good knowledge of lexical semantics. The first speaker provided initial input which was subsequently verified and extensively discussed with the expert speaker in order to ensure the coherence of the input and to avoid misunderstandings. We also made sure that native speakers received clear prior instructions on what is meant by a concept being lexicalized or not in a language. These instructions covered the following principles:

- *general vs specialized language:* we only consider terms that sound natural in general spoken language: thus, *sibling* is accepted but *nibling*, extremely rare and only known to specialists, is not;
- *restricted collocations vs free combinations:* fixed expressions that are frequently used are acceptable as lexicalizations (e.g. *little brother*), while expressions that are not felt as fixed and frequent should be considered as gaps (e.g. *female cousin*);
- *usage context:* speakers were encouraged to signal (in writing) stylistic, dialectal, or other constraints of usage for the words provided.

Native speakers were provided with the full concept list of every subdomain, with concepts described in English (all contributors were fluent English speakers), such as *"elder sister's child (as pronounced by a female speaker)"*. They were also given lexicalizations extracted from Wiktionary wherever available, as well as indications of lexical gaps that we automatically inferred. They were asked to validate these words and gaps, and also to provide lexicalizations and flag gaps for any concept not covered by Wiktionary. In the case of an incorrect word, they either had to provide a correct word or had to indicate it as a gap. In the case of an incorrect gap, they had to provide the appropriate word. Language-specific words extracted from Wiktionary for eight out of these ten languages were sufficient to infer

gaps. In Malayalam and Kannada, Wiktionary provided insufficient input and thus could not be used for gap inference. Ultimately, 165 words were retrieved by Wiktionary, 1,059 gaps were automatically inferred, and 230 words and 444 gaps were provided by native speakers as in Table 2.

Language	Wiktionary words	Inferred gaps	Expert words	Expert gaps
Arabic	22	126	6	36
English	16	134	16	20
French	20	129	16	29
Hindi	38	124	7	16
Hungarian	22	127	13	28
Italian	16	136	10	24
Kannada	1	0	63	128
Malayalam	3	0	60	131
Mongolian	12	144	23	7
Spanish	16	139	16	25
Total	165	1,059	230	444

Table 2: Statistics of the evaluation data.

Evaluation results

Our gap inference rules were conservative by design in order to favor precision, which is reflected in our evaluation results. Precision over inferred lexical gaps is very high, in the 99–100% range both across languages (Table 4) and subdomains (Table 5). Gap recall is 85.1%. False positive gaps only occurred in Hungarian and Mongolian. Hungarian *nagyszülő* "grandparent" was absent from Wiktionary, and was subsequently assumed as a lexical gap. As only one term үеэл 'cousin' was retrieved for Mongolian, our first gap inference rule assumed that all indirect descendants (of distance 2 or higher) were lexical gaps. Mongolian speakers informed us, however, that үеэл ах "elder male cousin" and үеэл эгч "elder female cousin" were widely used collocations.

Domain	Languages	Gaps	Cohen's Kappa
grandparents	19	246	0.94
grandchildren	8	135	0.98
siblings	2	20	0.90
uncle/aunt	17	335	0.75
Total	22	772	0.89

Table 3: For each subdomain, the agreement between, on the one hand, our Wiktionary-inferred and native-verified data and, on the other hand, typological evidence from Murdock. The second and third columns provide the data size used for evaluation, in terms of the number of overlapping languages and gaps.

The overlap between gaps signalled by Murdock and inferred by our method consisted of 772 gaps in 22 languages (i.e. merely 2% of the entire gap dataset, see Table 3). We computed the agreement between the two data sources and, using Cohen's Kappa, we obtained a score of 0.89, further implying that the gaps we inferred are high quality.

As for existing lexicalisations: the overall precision over Wiktionary words (after extraction) was 96.9%. The slight loss of precision is due to the appearance of rare French and Spanish terms in Wiktionary. Thus, our French evaluators noted that *adelphe* for "sibling" is a newly coined and mostly unknown term (e.g. it does not appear in the 1996 edition of the *Robert*). Moreover, *aïeul* and *aïeule* were identified as obsolete for designating "grandfather" and "grandmother". *Cadet* was judged not to mean "younger sibling", as stated by Wiktionary, but rather "younger brother." Likewise, the Spanish gender-neutral *hermane* "sibling" is a very rarely used neologism according to our evaluators. Apart from these examples, the terms retrieved from Wiktionary turned out to be of very high quality.

Recall on words is relatively low at 59.5%, which is a sign of Wiktionary incompleteness. Native speakers (other than Malayalam and Kannada) provided 107 new words and collocations (Table 2). In particular, many missing terms are expressed through restricted collocations in English, Spanish, Mongolian, and Hungarian. For example, 23 new inputs made by Mongolian speakers were all restricted collocations; e.g. ач хүү "son's son," нагац ах "maternal uncle." Other examples are English *elder sister*, Spanish *hermano menor* "younger brother" or *tío materno* "maternal uncle," Hungarian *fiúunoka* "grandson" or *nagytestvér* "elder sibling." In addition, our French speakers identified some words or morphological alternations missing from Wiktionary, e.g. the colloquial French *tata* "aunt," *papi* "grandfather," but also *aîné* "elder brother".

Languages	Words			Gaps		
	P	R	F_1	P	R	F_1
Arabic	100.0	78.6	88.0	100.0	77.8	87.5
English	100.0	50.0	66.7	100.0	87.0	93.0
French	80.0	50.0	61.5	100.0	81.6	89.9
Hindi	100.0	84.4	91.5	100.0	88.6	94.0
Hungarian	100.0	62.9	77.2	99.2	81.8	89.7
Italian	100.0	61.5	76.2	100.0	85.0	91.9
Mongolian	100.0	34.3	51.1	98.6	95.3	96.9
Spanish	93.8	48.4	63.9	100	84.8	91.8
Total	96.9	59.5	73.7	99.7	85.1	91.8

Table 4: Native speaker evaluation of words and lexical gaps by language.

Subdomains	Words			Gaps		
	P	R	F_1	P	R	F_1
grandparents	93.3	66.7	77.8	99.0	80.5	88.8
grandchildren	100.0	71.4	83.3	100.0	85.3	92.1
siblings	91.7	56.9	70.2	100.0	88.3	93.8
uncle/aunt	100.0	47.3	64.2	100.0	83.2	90.8
nephew/niece	100.0	63.6	77.8	100.0	85.3	92.1
cousins	100.0	60.4	75.3	99.4	86.1	92.3
Total	96.9	59.5	73.7	99.7	85.1	91.8

Table 5: Native speaker evaluation of words and lexical gaps by domain.

5 The Published Resource

Based on evaluation results, we considered the precision of our gap inference rules to be high enough to be applied to the final dataset without any modification. They were re-

Domain	Concepts	Relations	Murdock		Wiktionary + Expert			Total	
			Languages	Gaps	Languages	Words	Gaps	Languages	Gaps
grandparents	19	31	459	6,137	99	391	1,280	539	7,171
grandchildren	27	55	183	3,763	72	202	1,457	247	5,049
siblings	21	33	162	1,762	145	498	2,109	304	3,851
uncles/aunts	31	51	559	15,188	83	312	1,650	625	16,503
nephews/nieces	33	47	n/a	n/a	65	214	1,606	65	1,606
cousins	67	130	n/a	n/a	60	294	3,190	60	3,190
Total	198	347	561	26,850	168	1,911	11,292	699	37,370

Table 6: Statistics of the lexical gap resource.

File	Descrption	Columns
Concepts	lexical kinship concepts in interlingua	*subdomain*, *concept label*, *description*, *provenance*
Relations	hypernymy relations across concepts	*subdomain*, *hypernym concept label*, *hyponym concept label*
Words	lexicalizations in supported languages	*subdomain*, *concept label*, *lang name*, *ISO code*, *term*, *provenance*
Gaps	lexical gaps in supported languages	*subdomain*, *concept label*, *lang name*, *ISO code*, *evidence*

Table 7: Tab-separated files composing the kinship resource and their attributes.

run over the lexicalizations that were manually corrected based on native speaker input.

Statistics on the final resource obtained are provided in Table 6. The resource, as in Figure 1, is organized into a lexico-semantic interlingua layer and a layer of language-specific lexicons. The interlingua represents the six kinship subdomains through a total of 198 concepts and 347 hypernymy relations. In the lexicon layer, we automatically inferred 37,370 lexical gaps in 699 languages. 1,911 words were retrieved from Wiktionary and native speaker inputs. We inferred 26,850 gaps in 561 languages from Murdock data and 11,292 gaps in 168 languages from Wiktionary and native speaker inputs.

The resource is described online[2] and is also directly available for download.[3] It is distributed as four tab-separated text files, the structure of which is described in Table 7. The *concept label* column holds formal, structured representations of kinship concepts, such as `x;El;Br;Ch` meaning *elder brother's child as pronounced by a male speaker* (the last attribute is indicated by `x`). The *provenance* and *evidence* columns, in turn, provide the origin of the information, which can be: a reference to the typology research data, a reference to a lexical data source (e.g. Wiktionary), or *"native speaker."*

The kinship database has also been imported into the *Universal Knowledge Core* (UKC)[4], a multilingual lexical database of more than 1,000 languages (Giunchiglia et al., 2017; Giunchiglia et al., 2018). The focus of the UKC is language diversity: it is capable of representing lexical gaps and its online version is equipped with interactive visualisation tools that allow the browsing of kinship terms and gaps by subdomain in all supported languages (Bella et al., 2022), as shown in Figure 2.

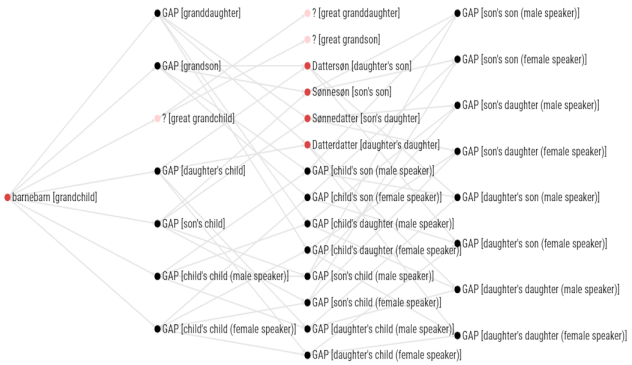

Figure 2: Interactive browser tool, showing lexicalizations and gaps for the grandchildren subdomain in Danish.

6 Application in Machine Translation

We believe that resources on lexical gaps have a high potential in improving existing cross-lingual applications, for example as pivots in multilingual representation learning, as seeds in cross-lingual model transfer, and also as part of multilingual word embeddings. In this section, we demonstrate how our resource can be used to improve and evaluate machine translation systems.

The Gap Problem in Machine Translation

Lexico-semantic diversity makes both human and machine translation challenging. Due to the lack of a concise translation equivalent for a given word or expression, translators need to find a term without excessively corrupting the original meaning. The following cases from the kinship domain demonstrate the challenge.

The sentences below are real machine translation (MT) outputs for the English sentence *My brother is three years younger than me* in five languages:[5]

[2] `http://ukc.disi.unitn.it/index.php/kinship/`

[3] `https://github.com/kbatsuren/KinDiv`

[4] `http://ukc.datascientia.eu`

[5] Obtained from Google Translate on 20 December 2021.

Target lang	Target sentence
Hungarian	*A bátyám három évvel fiatalabb nálam.
Japanese	*私の兄は私より3歳年下ですす.
Korean	*형은 나보다 세 살 아래다.
Mongolian	*Ах маань надаас гурван насаар дүү.
Russian	Мой брат младше меня на три года.

The only correct output is in Russian, due to the English *brother* having the equivalent брат. On the other hand, *brother* is a lexical gap in Hungarian, Japanese, Korean, and Mongolian, as these languages all express the relationship with more specific words. A corpus-statistics-based translation approach leads to a severe meaning-level mistake, as the Hungarian *bátyám*, the Japanese 兄, and the Mongolian ах all mean 'elder brother,' leading to the nonsensical result of *My elder brother is three years younger than me.* The Korean case is even worse as, besides assuming an elder brother, it also implies that the speaker is necessarily male.

In case of a non-existent equivalent, a semantically informed translation method can choose a broader term (e.g. *sibling*) instead of a narrower one, achieving approximation at the expense of slight information loss, less critical than inadvertently injecting false information. This works for Hungarian (*testvér*) but not for Mongolian that has no equivalent for *sibling* either, and where an even broader term such as *relative* may not sound right. In such cases, the appropriate narrower term (between ах "elder brother" or эрэгтэй дүү "younger brother") could, in this particular case, be inferred from the sentence context by a human translator or a sophisticated automated method.

The explicit and formal representation of untranslatability, as offered by our kinship resource, can be exploited to improve translation systems, but also to evaluate the semantic performance of existing systems on challenging tasks. In the rest of the section we illustrate the evaluation scenario through an experimental case study.

Machine Translation Evaluation Method

We describe a *semantic* evaluation measure and method for MT systems, designed to capture meaning-level translation mistakes that conventional metrics (e.g. the BLEU score) are known not to address adequately (Wu et al., 2016). The method consists of:

1. building a semantically annotated benchmark corpus of hard-to-translate sentences;
2. translating the sentences into a pre-defined set of target languages using an automated MT system;
3. measuring the *semantic distance* between key terms in the original and translated sentences.

A formal lexical gap resource, such as the one presented in this paper, is used in step 1, for the construction and annotation of sentences, and also in step 3 for the computation of semantic distances, based on the *least common subsumer distance* between the original and the translated term meaning in the interlingual concept hierarchy.

As a case study on step 1, we built a benchmark corpus of 50 English sentences that contain kinship terms, by adapting sentences from the British National Corpus (Burnard, 1995). The sentences contain 7–9 representative terms from each kinship subdomain: each such term appearing in a sentence was annotated by its meaning (in terms of the corresponding interlingual concept). Due to the well-known pervasiveness of lexical diversity among kinship terms, we consider this corpus as an effective meaning-level evaluation set of hard-to-translate sentences.

In step 2, we translated the 50 sentences into five languages: Hungarian, Japanese, Korean, Mongolian, and Russian, using Google Translate.

In step 3, we measured the semantic distance between target words and translation outputs. First, we automatically disambiguated output words (e.g. the Mongolian ах was disambiguated as "elder brother", formally represented as `El;Br`). Then we computed the semantic distance between this and the gold standard annotation. Figure 1 shows that the Mongolian ах and эрэгтэй дүү are connected through the concept of "brother." From "elder brother" to "younger brother" the distance is two hops, so the semantic distance amounts to 2.

Machine Translation Experiment Results

Table 8 shows for each language pair the number of gaps, the average semantic distance over gap-containing sentences, and the average semantic distance over all sentences. Machine translation of kinship terms turned out to be much more robust from English into Russian than into Japanese, Korean, Hungarian, or Mongolian, and a likely explanation for that is the fewer gaps encountered in Russian than the others.

On the whole, one can also observe that the gap-based distances are higher than the overall distances, which proves that handling gaps is indeed a weak point of current MT techniques that is likely worth addressing via dedicated solutions.

Language pair	Gaps	Sem.dist (gaps)	Sem.dist (all)
English–Russian	6	1.00	0.34
English–Japanese	13	1.06	0.38
English–Korean	12	1.58	0.90
English–Hungarian	19	1.31	1.06
English–Mongolian	12	1.33	1.12

Table 8: Semantic evaluation of Google Translate from English towards five languages.

7 Related Work

The large-scale multilingual lexical databases that exist today have been, by and large, produced and used by two distinct communities of researchers, namely historical and computational linguists. The former community has produced the *Intercontinental Dictionary Series* (IDS) (Key and Comrie, 2015), the *Automated Similarity Judgement Program* (ASJP) (Wichmann et al., 2013), *CLICS* (List et al., 2017), *CLDF* (Forkel et al., 2018), and the *World Loanword Database* (WOLD) (Haspelmath and Tadmor, 2009). These resources typically consist of phonemic descriptions of words or transliterations, as modern orthographies are irrelevant to both comparative and historical linguistics. This

characteristic, however, makes these resources difficult to use in computational applications that target contemporary written language.

The computational linguist and NLP communities, on the other hand, rely on resources derived from and describing contemporary written language. Formal, computer-processable lexical databases, such as BabelNet (Navigli and Ponzetto, 2012), the Open Multilingual Wordnet (OMW) (Bond and Foster, 2013), or Concepticon (List et al., 2016), however, focus on representing sameness, i.e. word meanings shared across languages, and do not explicitly indicate untranslatability. In BabelNet and OMW, language-specific word meanings either are left out or are mapped to other languages in an approximative manner. Such inaccuracies lead to a Western-centric bias as the word meanings that are correctly mapped belong to dominant languages such as the English Princeton WordNet (Miller, 1995).

We are aware of two efforts that address the formal representation and methodology of building diversity-aware lexical databases: MultiWordNet (MWN) (Pianta et al., 2002) and the announced second version of the Open Multilingual Wordnet (OMW2) (Bond et al., 2020). MWN has inbuilt support for representing lexical gaps and provides about 300 and 1,000 lexical gaps in Hebrew (Ordan and Wintner, 2007) and Italian (Bentivogli and Pianta, 2000), respectively. The recent paper (Bond et al., 2020) announced that OMW2 would be based on the *Collaborative InterLingual Index* (CILI), which envisions a collaborative method for defining language-specific concepts and gaps. We are not yet aware of the availability of an actual resource that would correspond to the theoretical diversity-aware representational abilities of the CILI. As also put forth in Bentivogli and Pianta (2000) and Ordan and Wintner (2007), identifying lexical gaps in a systematic manner is far from trivial. Let us take the example of the English word *cousin* which has no equivalent in Hindi and the Hindi word चचेरा भाई meaning *son of father's brother* which, in turn, has no concise equivalent in English. The Hindi gap corresponding to *cousin* can be identified as part of a expert-driven lexicon translation effort—called *expansion* in Fellbaum and Vossen (2012)—that is frequently used to produce lexicons for relatively lower-resourced languages. This approach was used, for example, to provide around 600 gaps in the Unified Scottish Gaelic Wordnet (Bella et al., 2020) and 79 gaps in Mongolian WordNet (Ganbold et al., 2014; Batsuren et al., 2019). This method, however, is effort-intensive and does not provide any gaps in the reverse direction. Traditional bilingual dictionaries that explicitly indicate untranslatability may be a good source of gaps, which corresponds to the *merge* method used in MWN. Such high-quality dictionaries, however, are not available for lots of languages.

Our approach is different from both of these: instead of the entire lexicon, it focuses on a single domain that is well-known to be cross-lingually diverse, such as kinship, colors, or body parts. Instead of human experts or existing lexical resources, it relies on evidence from linguistic typology. As a result, in the given domain it provides a much more exhaustive coverage both in terms of the number of gaps per language and in terms of the number of languages covered. Lastly but perhaps most importantly, ours is a predominantly top-down method that, thanks to knowledge from linguistic typology, is based on a prior conceptual understanding of the domain at hand (e.g. for the kinship domain, the analysis of 162 languages by Murdock (1970)). The result is that we are able to construct a hierarchy of interlingual lexico-semantic domain concepts with considerable precision, with language-specific lexicalisations (or the lack thereof) easily mappable to the hierarchy for lots of languages. In contrast, the bottom-up *expand* and *merge* methods, envisaged for the CILI and used in MWN, proceed by gradually extending an English-specific hierarchy, as new languages and new concepts are "encountered". This leads to the need for a (sometimes non-monotonic) reorganisation of the concept graph as the knowledge about a given domain evolves based on cross-lingual evidence. As shown in Figure 1, the one-by-one addition of Mongolian, Greenlandic, and Korean leads to profound changes in the interlingual representation.

8 Conclusion and Future Work

Our paper and the corresponding resource aim to address a gap in current multilingual lexicons and cross-lingual applications, namely the representation and exploitation of lexical diversity. We formally capture diversity through the notions of language-specific concepts and lexical gaps, and provide a systematic method to produce such data in a semi-automated manner. Our first large-scale effort focused on the domain of kinship terminology, well known to be particularly diverse across languages and cultures. The resulting machine-readable resource provides a wide coverage of domain concepts and languages (198 kinship concepts covered in 699 languages), and is freely available both for online browsing and download. In the future, we plan to apply the method presented in this paper to formalize new domains that are known to be diverse, such as colors, food, or visual objects (Giunchiglia and Bagchi, 2021). One such ongoing project concerns culturally specific concepts in the languages of India (Nair et al., 2022). Finally, we believe that resources such as ours provide essential information to lexically-focused cross-lingual applications, such as multilingual language models or cross-lingual transfer. We have presented one such application in the context of machine translation, but we plan to explore other application areas in the future. We also plan to extend our machine translation experiments to additional state-of-the-art MT systems.

9 References

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Arora, A., Farris, A., Gopalakrishnan, R., and Basu, S. (2021). Bhas. acitra: Visualising the dialect geography of south asia. *LChange'21*, 2021:51–57.

Arora, A., Farris, A., Basu, S., and Kolichala, S. (2022). Computational historical linguistics and language diversity in south asia. *arXiv preprint arXiv:2203.12524*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Batsuren, K., Bella, G., and Giunchiglia, F. (2021). A large and evolving cognate database. *Language Resources and Evaluation*, pages 1–25.

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*.

Bentivogli, L. and Pianta, E. (2000). Looking for lexical gaps. In *Proceedings of the ninth EURALEX International Congress*, pages 8–12. Stuttgart: Universität Stuttgart.

Bjerva, J., Salesky, E., Mielke, S. J., Chaudhary, A., Giuseppe, C., Ponti, E. M., Vylomova, E., Cotterell, R., and Augenstein, I. (2020). Sigtyp 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11.

Brown, C. H. (1976). General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature 1. *American ethnologist*, 3(3):400–424.

Catford, J. C. (1978). *A linguistic theory of translation*. Oxford University Press,.

Cotterell, R., Kirov, C., Hulden, M., and Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Fellbaum, C. and Vossen, P. (2012). Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10.

Gast, V. and Koptjevskaja-Tamm, M. (2018). The areal factor in lexical typology. In *Aspects of linguistic variation*, pages 43–82. De Gruyter Mouton.

Georgakopoulos, T., Grossman, E., Nikolaev, D., and Polis, S. (2021). Universal and macro-areal patterns in the lexicon. *Linguistic Typology*.

Giunchiglia, F. and Bagchi, M. (2021). Classifying concepts via visual properties. *arXiv preprint arXiv:2105.09422*.

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

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*, pages 18–24.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2016). A representation learning framework for multi-source transfer parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–325.

Joo, I. (2021). The etymology of korean ssal 'uncooked grain'and pap 'cooked grain'. *Cahiers de Linguistique Asie Orientale*, 50(1):94–110.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., and Cook, R. (2009). *The world color survey*. CSLI Publications Stanford, CA.

Kemp, C. and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

Koch, P. (2001). Lexical typology from a cognitive and linguistic point of view. inm. haspelmath, e. koenig, w. oesterreicher, & w. raible (eds.), language typology and language universals: An international handbook (pp. 1142–1178).

Koptjevskaja-Tamm, M., Rakhilina, E., and Vanhove, M. (2015). The semantics of lexical typology. In *The Routledge handbook of semantics*, pages 450–470. Routledge.

Levinson, S. C. and Evans, N. (2010). Time for a sea-change in linguistics: Response to comments on 'the myth of language universals'. *Lingua*, 120(12):2733–2758.

Levshina, N. (2021). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*.

McCarthy, A. D., Wu, W., Mueller, A., Watson, W., and Yarowsky, D. (2019). Modeling color terminology across thousands of languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2241–2250.

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, 9(2):165–208.

Nair, N. C., Velayuthan, R. S., Chandrashekar, Y., Bella, G., and Giunchiglia, F. (2022). IndoUKC: a Concept-Centered Indian Multilingual Lexical Resource. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

Padó, S. and Lapata, M. (2005). Cross-linguistic projection of role-semantic information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 859–866.

Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., and Cotterell, R. (2021). A surprisal–duration trade-off across and within the world's languages. *arXiv preprint arXiv:2109.15000*.

Ponti, E. M., O'horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., and Korhonen, A. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A. W., and Eisner, J. (2020). A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546.

Snyder, B. and Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of acl-08: hlt*, pages 737–745.

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)*.

Viberg, Å. (2014). The verbs of perception: A typological study. In *Explanations for language universals*, pages 123–162. De Gruyter Mouton.

Wierzbicka, A. (2008). Why there are no 'colour universals' in language and thought. *Journal of the Royal Anthropological Institute*, 14(2):407–425.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

10 Language Resource References

Batsuren, K., Ganbold, A., Chagnaa, A., and Giunchiglia, F. (2019). Building the mongolian wordnet. In *Proceedings of the 10th global WordNet conference*, pages 238–244.

Bella, G., McNeill, F., Gorman, R., Donnaíle, C. Ó., MacDonald, K., Chandrashekar, Y., Freihat, A. A., and Giunchiglia, F. (2020). A major wordnet for a minority language: Scottish gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Bond, F., da Costa, L. M., Goodman, M. W., McCrae, J. P., and Lohk, A. (2020). Some issues with building a multilingual wordnet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3189–3197.

Burnard, L. (1995). *Users reference guide for the British National Corpus*. Oxford University Computing Services.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ganbold, A., Farazi, F., Reyad, M., Nyamdavaa, O., and Giunchiglia, F. (2014). An experiment in managing language diversity across cultures. In *The Sixth International Conference on Information, Process, and Knowledge Management*. Citeseer.

Martin Haspelmath et al., editors. (2009). *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mary Ritchie Key et al., editors. (2015). *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., et al. (2016). D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, 11(7):e0158391.

List, J.-M., Cysouw, M., and Forkel, R. (2016). Concepticon: A resource for the linking of concept lists.

List, M., Greenhill, S., Anderson, C., Mayer, T., Tresoldi, T., and Forkel, R. (2017). Database of cross-linguistic colexifications.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Ordan, N. and Wintner, S. (2007). Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Wichmann, S., Muller, A., Velupillai, V., Brown, C. H., Holman, E. W., Brown, P., Sauppe, S., Belyaev, O., Urban, M., and Molochieva, Z. (2013). The asjp database (version 16). *Leipzig: Max Planck Institute for Evolutionary Anthropology*.