Check for updates

Lexical diversity in kinship across languages and dialects

Hadi Khalilia[1,2]*, Gábor Bella[3], Abed Alhakim Freihat[1],
Shandy Darma[1] and Fausto Giunchiglia[1]

[1]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy,
[2]Department of Computer Science, Palestine Technical University – Kadoorie, Tulkarm, Palestine,
[3]Lab-STICC CNRS UMR 628, IMT Atlantique, Brest, France

Languages are known to describe the world in diverse ways. Across lexicons, diversity is pervasive, appearing through phenomena such as lexical gaps and untranslatability. However, in computational resources, such as multilingual lexical databases, diversity is hardly ever represented. In this paper, we introduce a method to enrich computational lexicons with content relating to linguistic diversity. The method is verified through two large-scale case studies on kinship terminology, a domain known to be diverse across languages and cultures: one case study deals with seven Arabic dialects, while the other one with three Indonesian languages. Our results, made available as browseable and downloadable computational resources, extend prior linguistics research on kinship terminology, and provide insight into the extent of diversity even within linguistically and culturally close communities.

1 Introduction

The culture and the social structure of a community are reflected in the language spoken by its members. One of the most salient examples of this phenomenon is the worldwide diversity of terms used to describe family structures and relationships. While, thanks to studies such as Murdock (1970), kin terms around the globe are generally well-documented, many local variations—across dialects of a single language or across languages of a single country—have not yet been fully described or understood. For example, the term معزوزي *maazoozi* in the Algerian Arabic dialect, meaning *younger brother*, does not have any equivalent term in the Gulf Arabic dialect. In contrast, the Gulf word ابن العُود *ibn alood* meaning *elder brother* does not exist in Algerian, which instead uses the word سِيدي *siedi*.

Beyond a linguistic or anthropologic interest, the availability of digital resources on language diversity is also desirable from a computational perspective. Language processing applications need to be aware of such phenomena of diversity in order to provide high-quality results. For example, a machine translation system needs to tackle cases of lexical untranslatability where a word or expression in a source language has no equivalent in a given target language, and the choice of an approximate translation can change the meaning of an utterance. For example, for the English sentence *his cousin gave birth to a twin*, Google Translate provides the Arabic translation أنجب ابن عمه توأما *a'njaba ibna a'mihi tawaman* that means *His father's brother's son gave birth to a twin*. This syntactically correct yet unintended meaning of a male giving birth output is due to a *lexical gap*, i.e., a non-existent equivalent Arabic term for *cousin*. Such cases of *techno-linguistic bias*—where language technology provides better results *by design* in certain languages than in others—tend to remain hidden in monolingual resources but are revealed in multilingual settings (Bella et al., 2022a, 2023).

In recent years, there has been an increasing number of linguistic databases covering a large number of languages. These resources are usually aimed at quantitative studies for comparative linguistics, such as the classification of pain predicates (Reznikova et al., 2012), a semantic map of motion verbs (Wälchli and Cysouw, 2012), the modeling of color terminology (McCarthy et al., 2019), the CLICS database of cross-linguistic colexifications (Rzymski et al., 2020), DiACL (Diachronic Atlas of Comparative Linguistics), a database for ancient Indo-European languages spoken in Eurasia typology (Carling et al., 2018), or the Cross-Linguistic Database of Phonetic Transcription Systems (Anderson et al., 2018). Often, such databases use phonetic representations of lexical units or are limited to a few hundred or a few thousand core concepts, limiting their usability for the processing of contemporary written language. In our experience, most of the existing typology-informed NLP research is restricted to exploring language-specific morphosyntactic features and has ignored diversity within lexical resources (Batsuren et al., 2022). A notable exception is the Universal Knowledge Core, a massively multilingual lexical database that explicitly represents linguistic diversity and that we reuse in our work.

Our research is part of the *LiveLanguage* initiative, the overarching objective of which is to create, publish, and manage language resources that are "diversity-aware"—i.e., that reflect the viewpoints of multiple speaker communities—and that can be reused by multiple communities: linguists, cognitive scientists, AI engineers, language teachers and students (Bella et al., 2023). Contrary to mainstream exploitative practices, LiveLanguage aims to carry out its goals while empowering local speaker communities, giving them control over resources they help to produce (Helm et al., 2023). Involving human contributors and deciders from speaker communities is therefore a crucial part of our methodology.

In particular, the present paper focuses on diversity where it is less expected to appear: within dialects of the same language and within languages of the same country. Therefore, we describe a multidisciplinary study on the diversity of kin terms across seven Arabic dialects (Algerian, Egyptian, Tunisian, Gulf, Moroccan, Palestinian, and Syrian) and three languages from Indonesia (Indonesian, Javanese, and Banjarese). We consider kin terms as a domain particularly well-suited both for research on the methodology of collecting and producing diversity-aware linguistic data, and for comparative studies on diversity across languages.

Our paper aims to provide four contributions: (1) a general method for collecting multilingual lexical data from native speakers for a given domain (in our case the domain of kin terms), in a diversity-aware manner; (2) 223 kin terms and 1,619 lexical gaps collected in seven Arabic dialects and three Indonesian languages; (3) a qualitative and quantitative discussion of our results regarding the diversity observed across the dialects and languages covered; and (4) the publication of our results as an open, computer-processable dataset, as well as its integration into the Universal Knowledge Core multilingual database. Our starting point is state-of-the-art datasets on worldwide kinship terminology from ethnography (Murdock, 1970) and computational linguistics (Khishigsuren et al., 2022). Our data collection method is based on collaborative input from native speakers and language experts. Our results extend the state-of-the-art resources above with kin terms in languages and dialects not yet covered, as well as with 22 new kinship concepts not yet associated with other languages within those resources.

The structure of the paper is organized as follows. In Section 2, we give an overview of lexical typology and the phenomena of lexical untranslatability and lexical gaps with respect to the domain of kinship in particular. The Universal Knowledge Core resource is presented in Section 3. In Section 4, we describe our data collection method. Sections 5 and 6 introduce two case studies on Arabic dialects and Indonesian languages, respectively. Section 7 discusses previous studies related to our work. Finally, we provide conclusions in Section 8.

2 Untranslatability and lexical typology

Linguists understand translation from one language to another as a complex and multidimensional problem, ranging from multiple coexisting forms of meaning equivalence to untranslatability (Catford, 1965; Bella et al., 2022a). The diversity between cultures is a major cause for this problem appearing on several lexical-semantic levels. Some examples of the linguistic diversity are the richness of Toaripi vocabulary on the various forms of motion verbs describing walking around the beach like (isai) meaning "*go beachward*" and (kavai) meaning "*go inland with respect to the beach*", the language of the coastal Papua New Guinea country, the lack of vocabulary for the word meaning "*sailing*" in Mongolian, which is the language of a landlocked country, or the Arabic word سَنَّم meaning "*to ascend a camel's hump*".

The domain of kinship terms, which is the subject of our paper, is known to be extremely varied across languages, due to the different ways family structures are organized around the world. Matriarchal societies may describe certain female relatives with more detail, while strongly patriarchal ones are more descriptive with respect to male relatives. Arabic dialects, for instance, distinguish paternal and maternal brothers but also blood brothers, full brothers, and breastfeeding brothers. Thus, not only are kinship-related vocabularies "richer" or "poorer" across languages, they are also structured in different manners.

In this research, we focus on lexical untranslatability, which manifests most clearly through the lexical gap phenomenon when a word in a source language does not have a concise and precise translation in a given target language. Lexical gaps are often the linguistic manifestation of culturally or spatially defined specificities of a community of language speakers that cannot entirely be predicted or explained through systematic principles or recurrent patterns (Lehrer, 1970). Table 1 below presents this phenomenon for nine concepts representing sibling relationships from the kinship domain in eight languages.[1] One can observe that none of the

---

1   These nine concepts do not cover sibling terms exhaustively in all languages: for example, many Austronesian languages use different terms based on the gender of the speaker.

TABLE 1 Lexicalizations of nine meanings around the concept of (sibling) in eight languages.

Meaning	English	Japanese	Arabic	Italian	Indonesian	Hindi	Hungarian	Javanese
sibling	sibling	GAP	GAP	GAP	saudara	सहोदर	testvér	sedulur
elder sibling	GAP	GAP	GAP	GAP	kakak	GAP	nagytestvér	GAP
younger sibling	GAP	GAP	GAP	GAP	adik	GAP	kistestvér	adhi
brother	brother	GAP	أَخ	fratello	GAP	भैया	GAP	GAP
sister	sister	GAP	أُخْت	sorella	GAP	बहन	GAP	GAP
elder brother	GAP	あに	GAP	fratellone	abang	भैया	báty	kangmas
elder sister	GAP	あね	GAP	sorellona	GAP	दीदी	növér	mbakyu
younger brother	GAP	おとうと	GAP	fratellino	GAP	भाई	öcs	GAP
younger sister	GAP	いもうと	GAP	sorellina	GAP	बहन	húg	GAP

eight languages has concise lexicalizations for all nine concepts, yet each concept is lexicalized in at least one language. Such variations in lexicalization pose a problem for both machine and human translation: for instance, substituting a specific term instead of a broader one may result in injecting unintended meaning. In Javanese, at least four specific terms—(sedulur/*sibling*), (adhi/*younger sibling*), (kangmas/*elder brother*), and (Mbakyu/*elder sister*)—are used for expressing the sibling relationship, and accordingly, translating this sentence through Google Translate (*my sister is ten years older than me*) to Javanese gives this non-sensical sentence (*adhiku luwih tuwa sepuluh taun tinimbang aku*) meaning (*my younger sibling is ten years older than me*). This result is due to the lack of Javanese vocabulary for the word meaning (sister), and also lacks the term meaning "*younger sister*", so the machine translator uses (adhi) meaning "*younger sibling*," which finally produces the semantically absurd output.

Lexical typology is a field of linguistics that studies the diversity across languages according to the structural features of languages with respect to specific semantic fields (Plungyan, 2011). Different classical studies are conducted in this field on grammar and phonology, such as VoxClamantis V1.0–a large-scale corpus for phonetic typology (Salesky et al., 2020) and the structure of the space semantic field by identifying a set of semantic parameters and notions depending on the grammatical information of the field's constituents (Levinson and Wilkins, 2006). Other examples of such studies have been conducted on lexical-typological issues that appear across languages during translation, like the presence or absence of lexicalizations in languages. In these articles, authors focused on semantic fields that offer the richness of cross-lingual diversity: family relationships (Kemp and Regier, 2012), colors (Roberson et al., 2005), food (Bella et al., 2022b), body parts (Wierzbicka, 2007), putting and taking events (Kopecka and Narasimhan, 2012), cutting and breaking events (Majid et al., 2007), or cardinal direction terms (Arora et al., 2021). However, as mentioned in the introduction, only a few open datasets have been published in the scientific research area. These include the classification of kinship by Murdock (1970), which has been published in D-PLACE (Kirby et al., 2016). Part of Kay and Cook (2016)'s work on colors is published under the lexicon chapter of the World Atlas of Language Structures (WALS) (Dryer and

Haspelmath, 2013). Additionally, a color categorization dataset by McCarthy et al. (2019) is available on GitHub[2].

Digital lexicons have been increasingly used in lexical typology, enabling typologists to explore a broader range of languages and semantic domains. One noteworthy example is the KinDiv[3] lexicon (Khishigsuren et al., 2022), which encompasses 1,911 words and identifies 37,370 gaps within the domain of kinship, spanning 699 languages. In our current research, we extend our investigation into the kinship domain, specifically focusing on exploring linguistic diversity among Arabic dialects and Indonesian languages. Other examples include Viberg (1983)'s seminal study, which was conducted on perceptual terminology in 50 languages and has been expanded upon by Georgakopoulos et al. (2022) to cover 1,220 languages. Furthermore, the Kinbank database, recently introduced by Passmore et al. (2023), serves as a comprehensive repository of kinship terminology, encompassing more than 1,173 languages and offering a broad coverage of various kinship subdomains.

3 Universal Knowledge Core

This section describes the Universal Knowledge Core (UKC)[4], a large multilingual lexical database that we adopt for the production of diversity-aware datasets in this research (Giunchiglia et al., 2017). The use of the UKC is motivated by its ability to represent linguistic unity and diversity explicitly: conceptualizations shared across languages, word senses appearing only in certain languages, shared lexicalizations (e.g., cognates), as well as lexical gaps. The theoretical underpinnings of the lexical model of the UKC have been described in Giunchiglia et al. (2018) and in Bella et al. (2022b), and are illustrated in Figure 1.

The UKC is divided into a supra-lingual concept layer (as shown at the top of Figure 1) and the layer of individual lexicons (at the bottom of Figure 1). The concept layer includes hierarchies of concepts that represent lexical meaning shared across languages. Concepts are language-independent units and act as bridges across

FIGURE 1
Structural elements in the UKC lexical database.
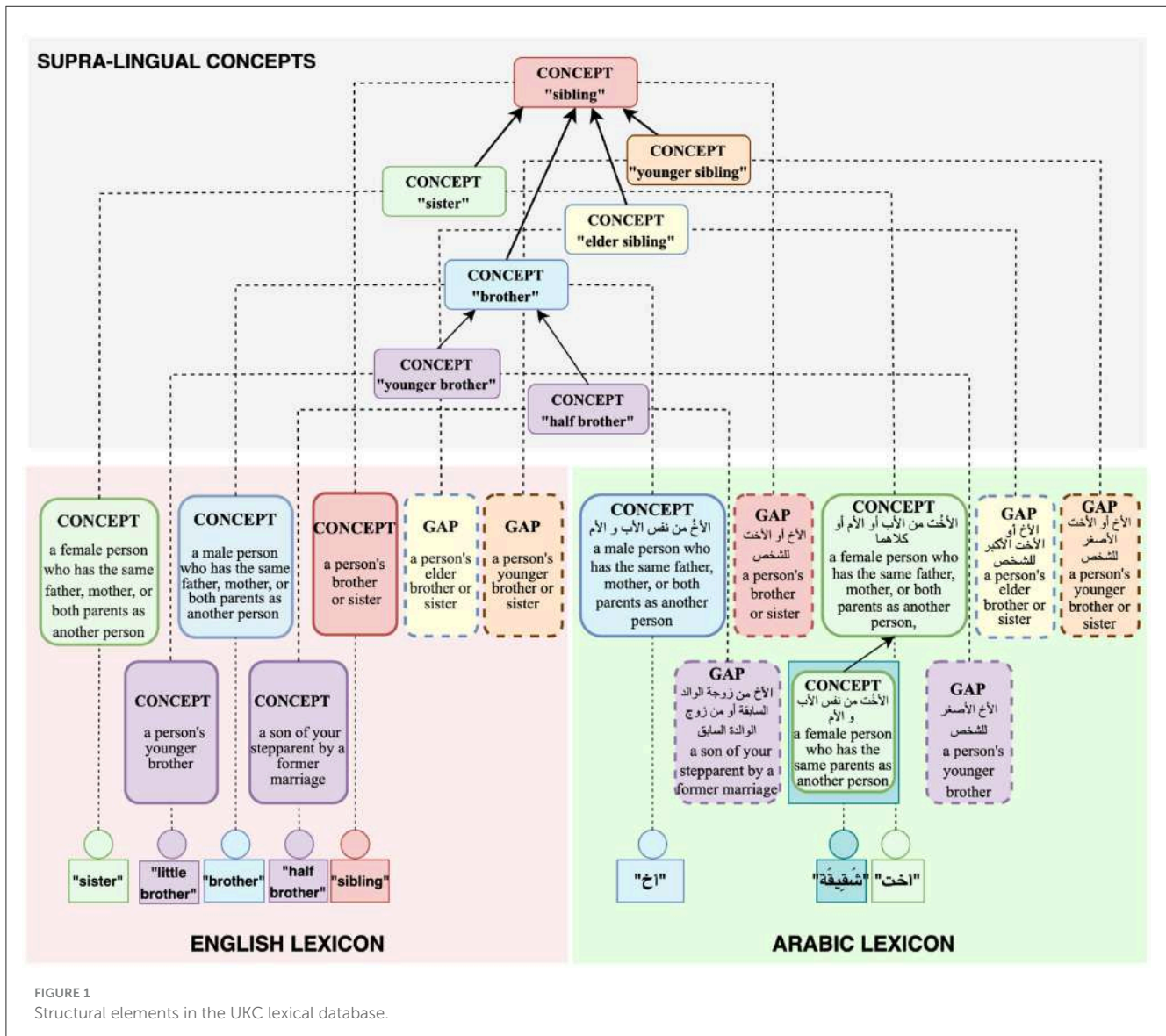
languages, and each one should be lexicalized by at least one language to be present in the concept layer. Supra-lingual concepts and their relations (e.g., hypernymy, meronymy) are in part derived from third-party resources such as Princeton WordNet (PWN) (Miller, 1995), and are in part proper to the UKC. In particular, the UKC contains an extensive formal conceptualization of kinship domain terms computed from the KinDiv database, spanning about 200 distinct concepts.[5] KinDiv itself is based on ethnographic evidence from 699 languages (Khishigsuren et al., 2022). While this existing hierarchy of kinship concepts does not fully cover all terms that appear in our study, it is the most complete one we are aware of, motivating our choice of the UKC as a platform for our research.

The lexicon layer consists of language-specific lexicons that provide lexicalizations for the concepts from the supra-lingual concept layer, while also asserting *lexical gaps* whenever

lexicalizations are known not to exist. Lexicons also provide term definitions as well as lexical relationships specific to the language, such as derivations, metonymy, or antonymy relations. Lexicons can also contain *language-specific concepts* that do not appear in the supra-lingual concept layer. For example, in Figure 1, the Arabic شَقِيقَة, meaning "*a female person who has the same father, mother, or both parents as another person*", is represented as a language-specific concept. The dual mechanism of defining lexical concepts either on the supra-lingual or on the language-specific level allows for the representation of differing worldviews that would be hard or impossible to reconcile into a single global concept graph. The richness of its lexicon-level linguistic knowledge makes the UKC unique among multilingual lexical databases and particularly suitable for our study.

As mentioned in Section 2, a lexical gap for a specific concept is present in a language if there is no concise equivalent word meaning for the concept in that language. For example, neither English nor Arabic has a word meaning *elder sibling*;

---

for such cases, the UKC provides evidence of meaning non-existence and untranslatability by representing lexical gaps inside lexicons, as shown in Figure 1. This information can be used by the NLP community to indicate the absence of equivalent words to downstream cross-lingual applications.

Beyond providing lexical relations between shared word meanings as other multilingual lexical databases do, the UKC also represents a richer set of lexical-semantic connections between language units in a lexicon. For example, the *antonym* lexical relation expresses that two senses are opposite in meaning. While the lexical-semantic relation, *similar-to*, is used to connect two concepts with similar meanings, and the *hypernym-of* connects parent meaning with its child. For instance, in Figure 1, the English (little brother) and (brother) are connected through a *hypernym-of* relationship. Such information can be used by the NLP community to indicate the concise equivalent language-specific word meaning to downstream cross-lingual applications, e.g., as the position of a language-specific meaning in a language hierarchy in a lexicon.

The UKC currently does not explicitly distinguish between languages and dialects: each vocabulary is a separate entity labeled with a standard three-letter ISO 639-3 code. When such a code is not available, the UKC uses a standard extension mechanism where three additional (not standardized) letters are added to the ISO code: e.g., for Syrian Arabic, the code `arb-syr` is used.

4 A methodology for building diversity-aware lexicons

This section presents the general method by which we collected and produced lexicalizations and gaps from native speakers and language experts. The same method presented below was employed in an independent manner for each Arabic dialect and Indonesian language covered by our study. The contents of this section aim to serve as a tried and tested recipe for gathering lexical data in a diversity-aware manner, that we intend to reuse in future lexicon development projects.

We exploit the UKC to import language-independent concepts (e.g., kinship concepts) to be used as an input dataset to our method and use its data representation model to formalize our data. We reuse an already broad and well-formalized hierarchy of 184 kinship concepts from the KinDiv database, which includes kinship terms and gaps in 699 languages. Data in KinDiv is based on the well-known results of Murdock (1970), as well as on lexicalizations retrieved from Wiktionary that we consider as an overall good-quality resource. In Khishigsuren et al. (2022), the accuracy of KinDiv was evaluated to be above 96%. One language expert per language provided this percentage, which represents the proportion of the number of words (or gaps) validated as correct to the total number of collected words (or gaps).

Our work extends KinDiv data by new concepts, lexicalizations, and lexical gaps in languages and dialects that are either not present in KinDiv or are incompletely covered. A lexical-semantic expert generates a contribution (kinship terms and gaps) task, then a group of native speakers collects contributions from a dialect (and a local language). After that, two steps for validating collected contributions: language experts evaluate collected lexical units and gaps of a dialect, and a lexical-semantic expert evaluates explored

kinship concepts (not existing in UKC). Additionally, resulting data (including gaps, words, and new concepts) is used to update and enrich UKC. So, gaps and words are merged into the lexicons of the UKC while new concepts are integrated with the (top) concept layer. A general view of the method is depicted in Figure 2.

Accordingly, the macro-steps of our methodology are as follows:

1. *Contribution task generation*: First, prepare the materials: the dataset of inputs to be examined and the architecture of the supra-lingual concept layer of each subdomain.
2. *Contribution collection*: The actual contribution effort is carried out by a native speaker in a local language or dialect.
3. *Lexicon-level validation*: Provided words and gaps are evaluated and corrected by a language expert.
4. *Concept-level validation*: New concepts and unclear contributions (i.e., words on the borderline) are verified by a lexico-semantic expert.

4.1 Contribution task generation

This section describes the material needed during the execution of the next steps of the methodology. Hence, two constituents must be prepared in this step as described below:

1. *Dataset of inputs*: Constructing the dataset of general word meanings is the first step of studying diversity across dialects and represents the inputs of the contribution collection phase. In this context, the UKC lexicon is employed to build a dataset, which contains several facilities that support retrieving categorized data from its interlingual shared meaning layer as introduced in Section 3. Moreover, typology datasets or other approaches can be used for that, such as the kinship dataset from Murdock (1970); or gathering data from online dictionaries using automatic methods, i.e., KinDiv retrieves some of its kinship terms from Wiktionary. The constructed dataset is a spreadsheet containing language-independent meanings from one semantic field. At the same time, its content is distributed into subdomains (sheets) for usability and simplicity in designing a concept hierarchy for each subdomain which is a helpful tool for lexical-gap exploration. One spreadsheet row is generated for each concept, containing the concept ID, the source concept definition in the standard language, another definition in English, as well as empty slots for inserting a lexical gap or a word with equivalent meaning, and the data provider's comments in a dialect or local language.
2. *Interlingual concept hierarchy*: Modeling the interlingual shared meaning space is essential to explore lexical gaps systematically. In this task, the UKC concept hierarchy is exploited. UKC is the only resource introducing a hierarchy of shared meanings across languages for each semantic field, such as kinship, colors, or food. Furthermore, UKC uses a hybrid linguistic-conceptual approach in modeling each domain. This approach adopts actual domain ontology and linguistic data from typological literature. For example, a fragment of the brotherhood hierarchy in the top layer of the UKC is shown in Figure 1. A native speaker can compare each examined concept from the spreadsheet with the hierarchy of its domain to extract additional knowledge

FIGURE 2
Methodology macro-steps and data sources.
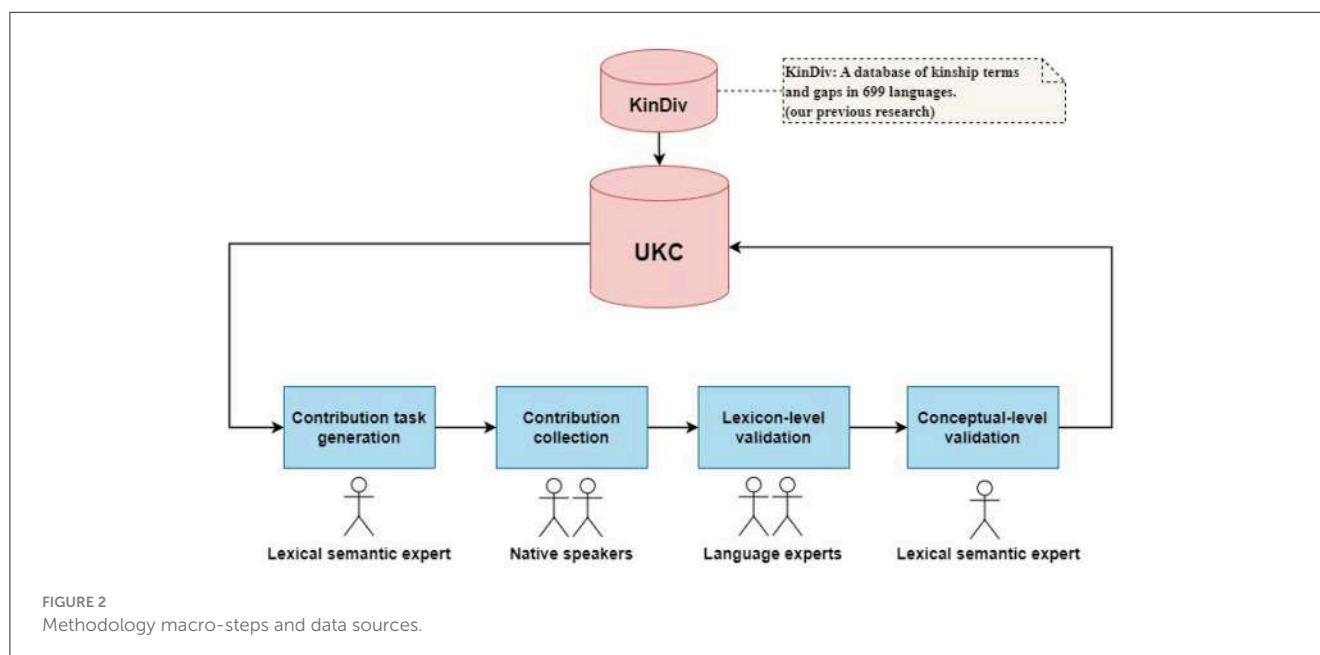
about its meaning based on a concept's position in the hierarchy, which helps to provide a concrete answer in terms of a gap or a lexical unit.

4.2 Contribution collection

Contributions from a local language or a dialect are provided by one native speaker who was born and educated (university level) within the speaker community. The following are the most notable instructions they are given:

1. They are given the authority to skip concepts, stop contributions, or leave a comment when they deem the terms are becoming too culture-specific and consequently need an exact answer.
2. They are asked to provide a lexicalization in a local language (or dialect) that gives meaning equal to the concept's meaning.
3. They are asked explicitly to identify lexical gaps where no local (or dialect) lexicalization exists.
4. Within a local language (or dialect) and a subdomain (e.g., cousins), they are asked to provide new concepts that did not exist in the list of inputs which is imported from the UKC by providing a word (lemma) and a clear description of its meaning.

The process of providing such contributions is depicted in two flowcharts; for instance, Figure 3 shows the flowchart of the candidate gap (on the left-hand side of the figure) and candidate equivalent word meaning (on the right-hand side of the figure) exploration; it starts identifying a standard language and a local language (or dialect) and providing a native speaker with a spreadsheet including a list of subdomain concepts (inputs). Then, a native speaker is asked to find a linguistic resource in the local language and use it to search for concepts (concept-by-concept) to confirm lexicalizations and gaps. He/she can use a linguistic resource in the search process as the following steps: searching in a well-known dictionary, then in Wiktionary—a large multilingual online lexicon after that in a typology dataset (if it is available),

and finally, using Google search (based on the count of search hits). More details about these steps are described in Section 5. The native speaker can rely on search results and the count of Google hits to give a more concrete answer on whether the concept in the standard language has a lexicalization or is a gap in the local language; such candidates are passed to the next phase- lexicon-level validation.

A new concept collection is a third contribution in this phase, where the steps of a candidate new concept exploration in a local language can be seen in Figure 4. A native speaker can examine the list of subdomain concepts and provide his/her (own) concepts with their definitions that he/she believes have not existed in the list. The same search steps in gap identification can be followed in this task. As shown in Figure 4, All candidate new concepts are passed to the two subsequent validation phases: lexicon- and conceptual level.

4.3 Lexicon-level validation

Our lexicon-level validation method formally and explicitly addresses individual gap identifications and their quality, as well as equivalent word meanings and new concepts. It allows a qualitative evaluation of the entire list of provided contributions through word-by-word and gap-by-gap in a loop between a native speaker and a validator. A word, a gap, or a new concept does not pass this validation until the native speaker provides the correct answer for each of them, as shown in the flowcharts in Figures 3, 4.

A language expert who is also a native speaker of the determined language (or dialect) will carry out this validation on a spreadsheet containing the data and results gathered in the previous step with two additional empty columns: the evaluation and lexicon-level validator's comment, producing the following information:

1. *Equivalent word meanings*: validate the correctness of all provided words in the local language (or dialect) by marking them up
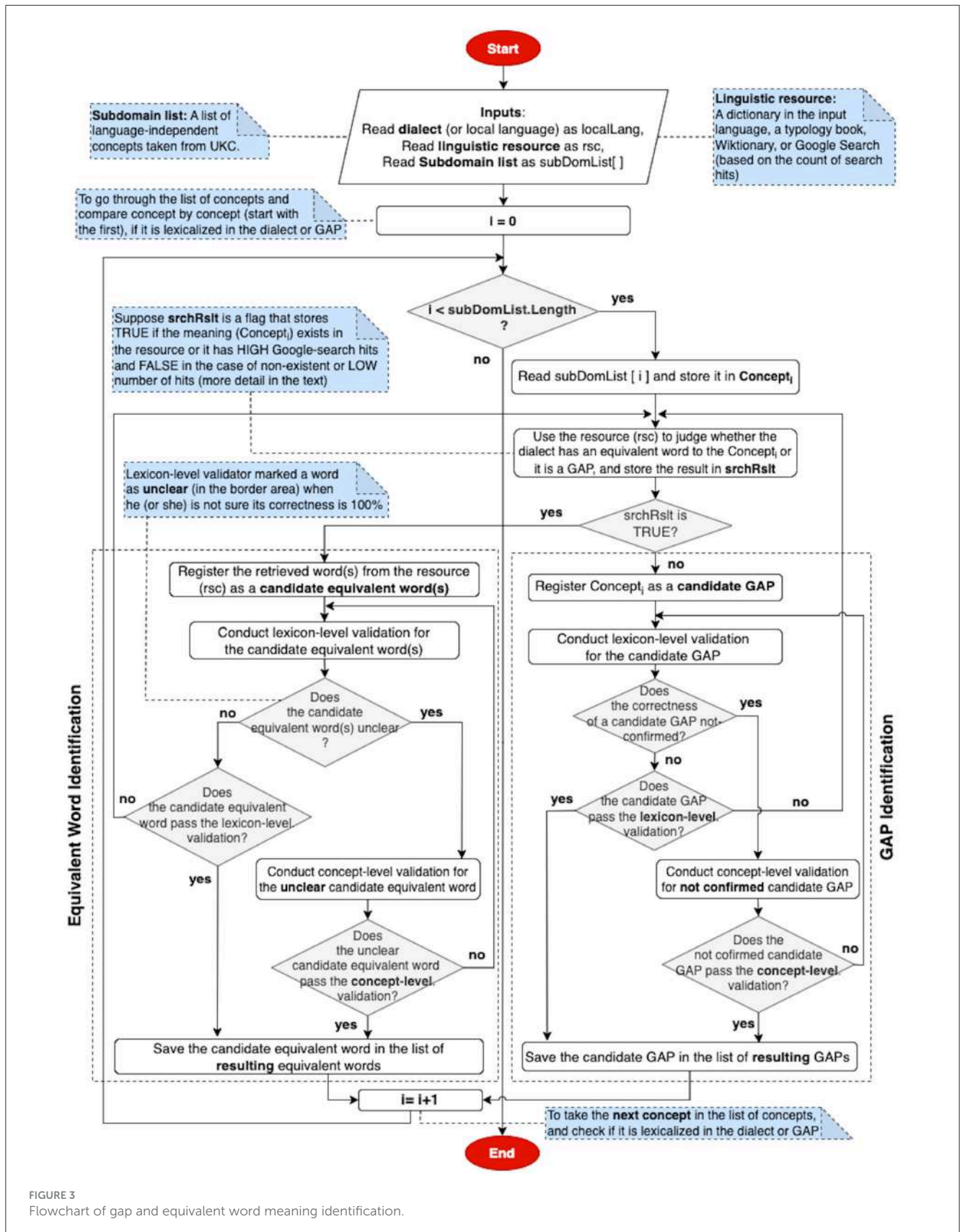
FIGURE 3
Flowchart of gap and equivalent word meaning identification.

as correct, incorrect, or unclear for borderline cases and by providing correct words or indicating them as lexical gaps for incorrect ones.

2. *Lexical gaps*: validate the word meanings marked as lexical gaps by the native speaker in the local language, either as confirmed gaps or as non-gaps due to an existing
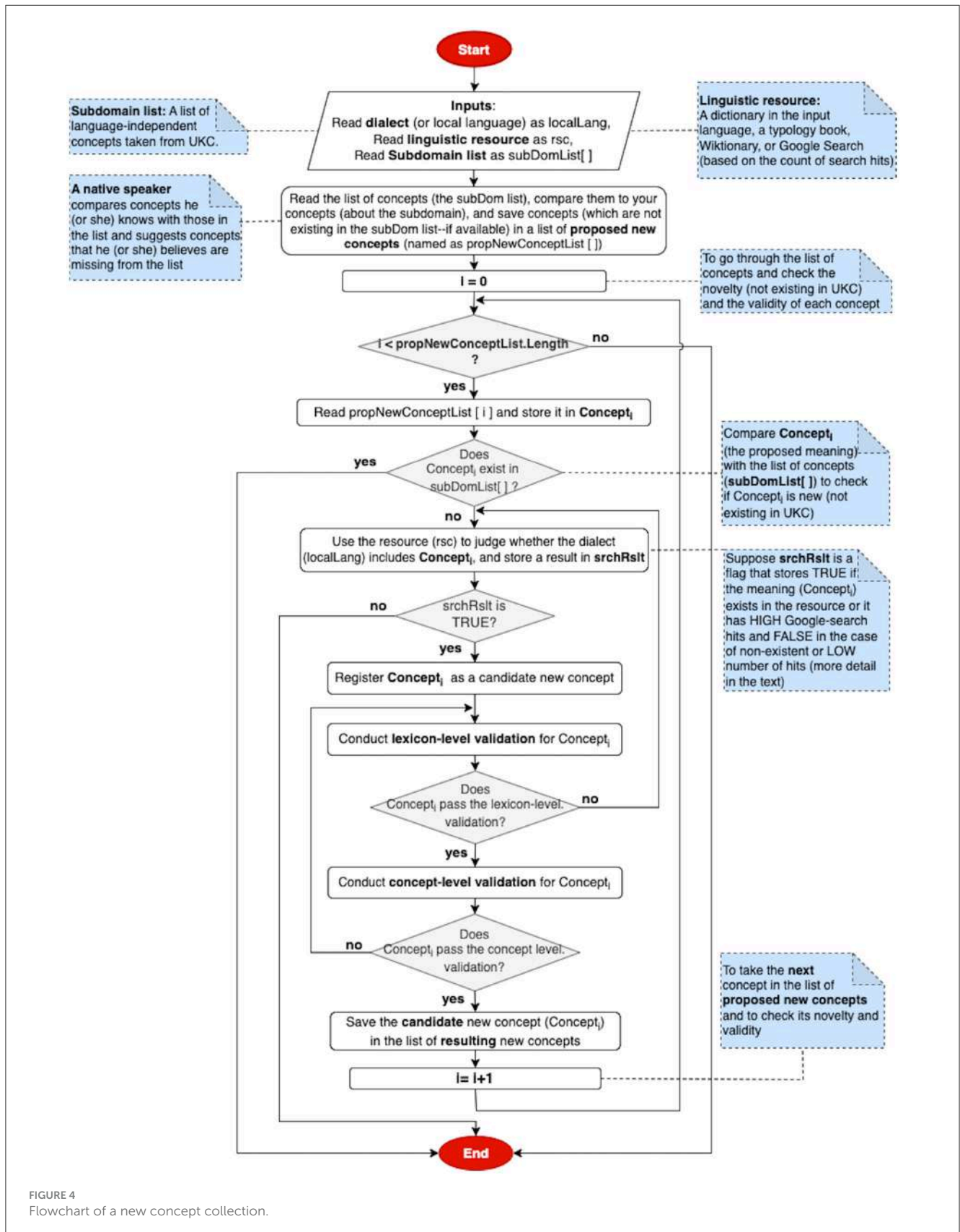
FIGURE 4
Flowchart of a new concept collection.

lexicalization in that language, which the validator needs to indicate.

3. *New concepts*: validate all proposed new word meanings in each subdomain by marking them up as correct, correct but not new

(in case the supposedly new concepts already existed in the list), or not accepted (in case another concept already existed in the list to express the meaning, or the validator does not consider it as a desirable suggestion for other reasons).

TABLE 2 The count of concepts in the input dataset.

Subdomains	Count of concepts
Grandparents	19
Grandchildren	27
Siblings	21
Uncle/aunt	27
Nephew/niece	33
Cousins	57
Total	184

TABLE 3 Count of Google search hits for cousin concepts in Arabic.

Concept	With/Without diacritics	Count of hits	
العمومة Paternal cousin	العُمومَةُ	1.94 M	3.04 M
	العمومة	1.1 M	
الخُؤولة Maternal cousin	الخُؤولَةُ	111 k	158 k
	الخؤولة	47 k	
ابن العم Son of father's brother	اِبْن العَمّ	84.8 M	93.96 M
	ابن العم	9.16 M	
بنت العم Daughter of father's brother	بِنْت العَمّ	8.43 M	83.13 M
	بنت العم	74.7 M	
ابن العمة Son of father's sister	اِبْن العَمَّة	12.5 M	131.5 M
	ابن العمة	119 M	
بنت العمة Daughter of father's sister	بِنْت العَمَّة	9 M	30.4 M
	بنت العمة	21.4 M	
ابن الخال Son of mother's brother	اِبْن الخَال	5.61 M	33.01 M
	ابن الخال	27.4 M	
بنت الخال Daughter of mother's brother	بِنْت الخَال	3.99 M	30.69 M
	بنت الخال	26.7 M	
ابن الخالة Son of mother's sister	اِبْن الخَالَة	12.5 M	16.59 M
	ابن الخالة	4.09 M	
بنت الخالة Daughter of mother's sister	بِنْت الخَالَة	11 M	16.67 M
	بنت الخالة	5.67 M	

Correct equivalent word meanings and gaps are integrated with the local language lexicon on the fly. Also, correct new concepts are passed to the next step to be validated at the concept level before merging them with the supra-lingual shared meaning layer. While in case the evaluation is an incorrect equivalent word or a gap, or not accepted new concept, the validator returns each of them with a comment describing the reason to the native speaker to review and address the problem; when the native speaker finishes revising them, then he/she returns the new version of a contribution to the validator. This cycle (native speaker's contribution—lexicon level validation) is still alive until the validator confirms the correctness of the contribution or skips it.

4.4 Concept-level validation

In this step, a lexical-semantic expert who is the manager of the UKC system verifies the new concepts and their quality as accept or reject to add them into the supra-lingual concept layer as well as addresses unclear words and non-confirmed gaps/non-gaps that are borderline cases. This validation is based on a discussion session with the language expert responsible for lexicon-level validation through concept-by-concept and case-by-case issue validation. A spreadsheet containing all new concepts and determined (words and gaps) to be examined is used. Columns of this sheet are the same columns in the previous step and two additional empty ones: the evaluation and concept-level validator's comment. The following tasks are used:

1. *New concepts*: Validate all proposed new concepts in each subdomain by marking them up as correct, correct but not new (in case the supposedly new concepts already existed in the UKC), or not accepted (in case another concept already existed in the UKC to express the meaning, or the validator does not consider the new concept as a desirable suggestion for any other reason).
2. *Unclear words*: Validate the correctness of unclear word cases considered in the border-area by the lexicon-level validator by marking them as correct or incorrect and writing a comment.

3. *Non-confirmed gaps/non-gaps*: Validate the word meanings that do not have confirmation as lexical gaps or non-gaps by providing a judgment with a comment.

Correct new concepts are imported into UKC by merging them with the supra-lingual conceptual layer. In contrast, not-accepted ones and those correct but not new are returned to the validator at the lexicon level, who may also return them with a comment describing the reason to the native speaker to address an included problem. In a new cycle, modified new concepts by the native speaker are transferred to this phase through the validator of lexicon-level; then, the validator at this level reviews the updates and decides whether to finish the revision cycle by accepting or rejecting the new concepts or issue a new one for more review, as shown in Figure 4. In addition, confirmed words and gaps output from this step are integrated with the language lexicon in the UKC, as shown in Figure 3.

TABLE 4 The count of the diversity items collected and identified in the Arabic dialects.

Dialects	Words	Gaps w/o new concepts	New concepts	Gaps considering new concepts
Algerian	28	156	10	165
Egyptian	32	152	19	152
Moroccan	22	162	10	169
Palestinian	23	161	14	166
Syrian	24	160	10	169
Tunisian	23	161	2	178
Gulf	28	156	14	169
Total	180	1,108	19	1,168

TABLE 5 Validator evaluation of words and lexical gaps by dialect.

Dialects	Correctness of native speaker contribution	
	Words (%)	Gaps (%)
Algerian	85.71	98.08
Egyptian	96.90	97.37
Moroccan	95.83	97.53
Palestinian	100	98.76
Syrian	91.67	95.00
Tunisian	95.65	98.14
Gulf	100	96.79
Average	95.11	97.38

5 Case study on diversity across Arabic dialects

This section demonstrates the use of the methodology described in Section 4 on kinship terminology from seven dialects of the Arabic language. Arabic is the official language of more than four hundred million native speakers in twenty-two countries in the Middle East and northern Africa. Classical Arabic or Modern Standard Arabic (MSA) refers to the standard form of the language used in academic writing, formal communication, classical poetry, and religious sermons (Elkateb et al., 2006). Surprisingly lexical diversity is manifested between Arabic dialects, evident in our study between seven of the twenty dialects spoken worldwide. The selected dialects are Egyptian, Moroccan, Tunisian, Algerian, Gulf, and South Levantine (two examples: Palestinian and Syrian). Let us take the example of the Gulf word الخَال العودْ meaning "*mother's elder brother*," which has no equivalent in South Levantine or Moroccan; instead, they use the more general word الخَال meaning "*mother's brother*," which can be used for both meanings "*mother's younger brother*" or "*mother's elder brother*". In this paper, we perform an experiment on the Arabic dialects to capture their diversity in the kinship domain. The resulting dataset with dialect-specific kinship terms will be integrated with an instance of the Universal Knowledge

Core for Arabic (Arabic UKC)[6] ongoing project, which is the first diversity-aware lexical resource for Arabic dialects so far.

5.1 Experiment setup

As mentioned in Section 3, the UKC resource is our data source in building the input dataset of kinship-independent language concepts and formalizing such concepts and new word meanings (not existing in the inputs) explored in this experiment. For example, the brotherhood hierarchy is shown in the top layer of the UKC in Figure 1. In this study, contributions are provided by seven native speakers (one per Arabic dialect). Regarding the contributors' socio-linguistic background, each has at least a master's degree and was born and educated, at least up to high school level, within the native speaker community. The participants' linguistic backgrounds are presented below:

1. *Participant 1*: a native Algerian speaker with good command of English.
2. *Participant 2*: a native Egyptian speaker with good command of English.
3. *Participant 3*: a native Tunisian speaker with good command of English and French.
4. *Participant 4*: a native Gulf speaker with good command of English and Arabic-Palestinian.
5. *Participant 5*: a native Moroccan speaker with good command of English and Italian.
6. *Participant 6*: a native Palestinian speaker with good command of Arabic-Syrian and English.
7. *Participant 7*: a native Syrian speaker with good command of English.

Seven experiments (one for each dialect) are performed to explore lexical units and gaps using our method. In each experiment, a spreadsheet of kinship concepts is imported from the UKC (as the source, they were computed from the KinDiv database), which serves as an input dataset to the contribution (diversity-aspects) collection step. These kinship domain concepts are language-independent units representing lexical meaning shared across 699

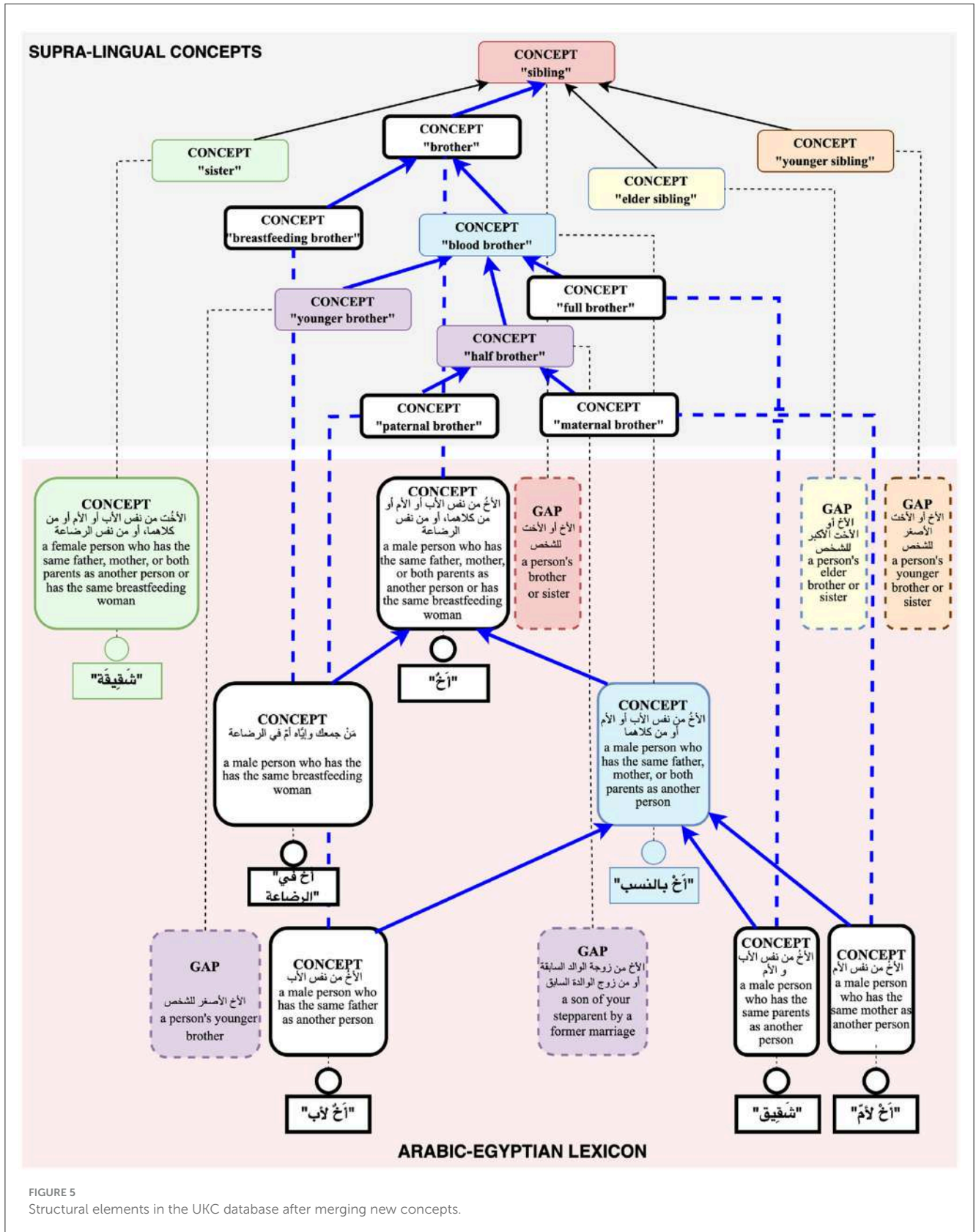6   http://arabic.ukc.datascientia.eu/concept

FIGURE 5
Structural elements in the UKC database after merging new concepts.

languages and spanning 184 distinct concepts. UKC categorizes kinship concepts into six groups; each one contains a distinct subset of concepts sharing a common kinship type meaning

called a subdomain, for example, sibling and cousin subdomains. The spreadsheet (the dataset) consists of six sheets, and each one represents a kinship subdomain. See Table 2, which shows
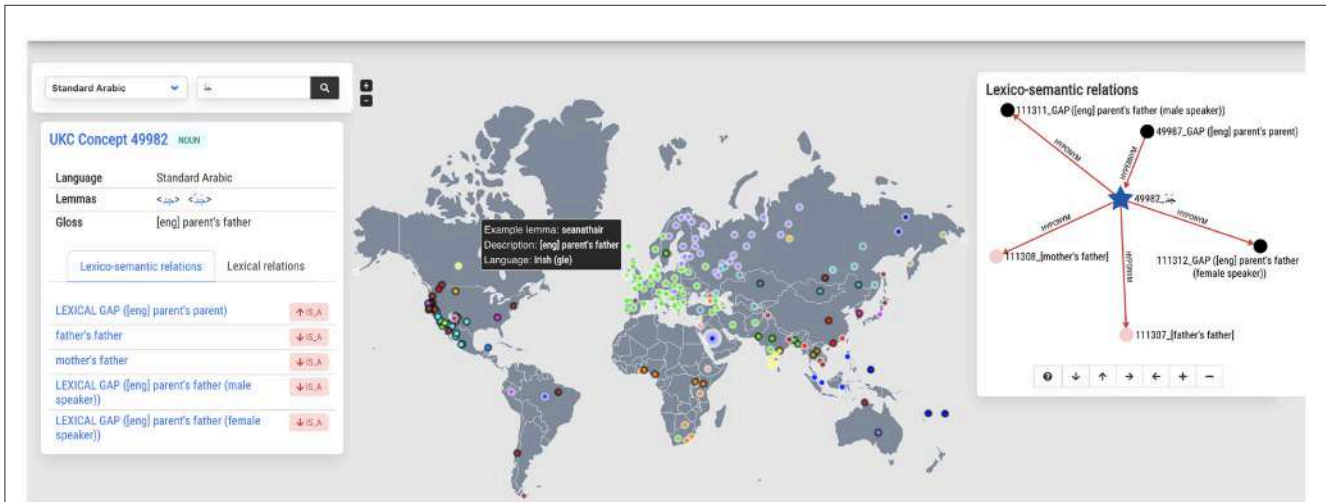
FIGURE 6

Exploring the concept of جَدّ as lexicalized in the Arabic language **(left)**, in the world **(middle)**, and as part of the shared concept hierarchy **(right)**.
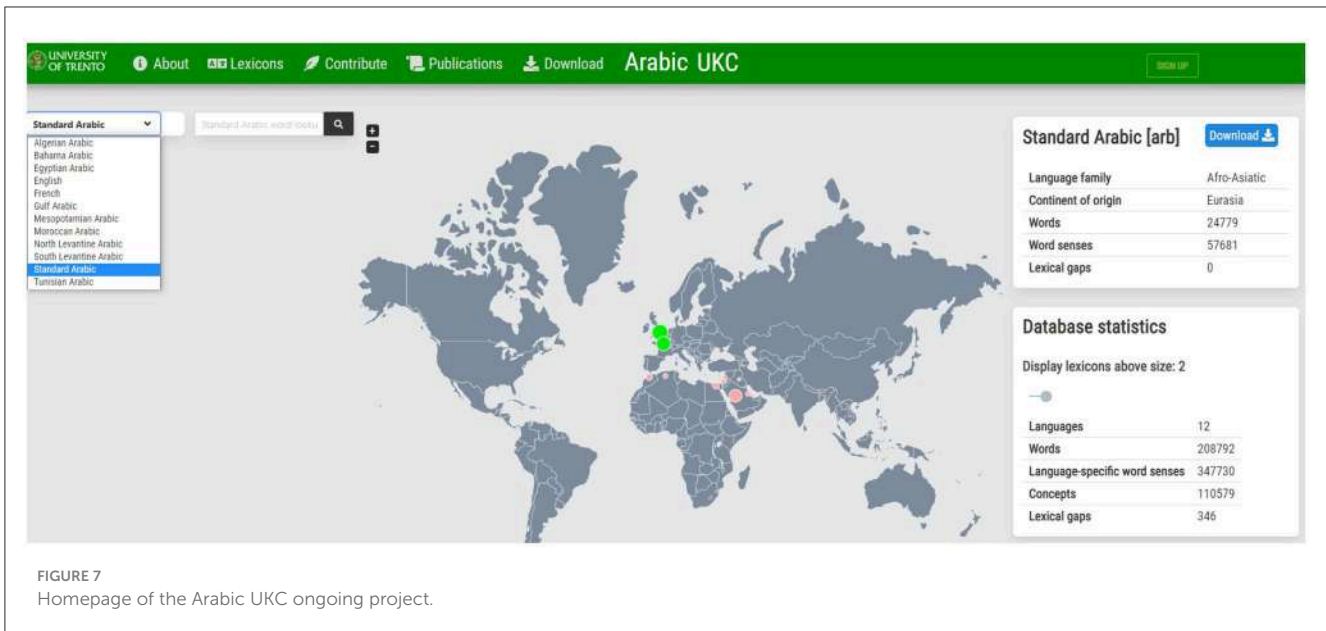

FIGURE 7

Homepage of the Arabic UKC ongoing project.

the subdomain names and the count of containing concepts per subdomain of the dataset.

In the contribution collection, a native speaker answers by filling a lexical unit or gap in a row empty slot specified for each concept. Linguistic resources and Google Search are used to provide answers as precise as possible. For example, the المعاني Almaany dictionary[7], Wiktionary[8], and the *Fiqh AlArabiyya* typology book (Muttaqin, 2009) are employed in sequential steps to give a judgment on cousin words in Syrian. Additionally, counting the number of hits returned by the Google search engine is another helpful indicator, where a high count of hits indicates a searching word (i.e., ابن العمة meaning "*son of father's sister*," has 131.5 million hits) is a lexical unit in Syrian. In contrast, a low count indicates a lexical gap; for example, الخؤولة

———

7   http://www.almaany.com/thesaurus.php

8   http://ar.wiktionary.org

meaning "*maternal cousin*," has 158 thousand hits. Google hits of other cousin terms are shown in Table 3. Since Arabic words can be written and read with or without diacritics (i.e., "*fatha*" above a letter or "*kassra*" under it), thus, each word is typed in two forms. Note that the content of this matrix cannot be considered the only criterion for gap exploration because word hits may contain a count of other hits resulting from searching in other Arabic dialects for the same word.

5.2 Experiment results

The overall contribution collection effort resulted in 180 words, 1,108 lexical gaps, and 19 new concepts identified, formalized, and collected. Detailed statistics about the collected gaps and words are shown in Table 4. New concepts were identified in three
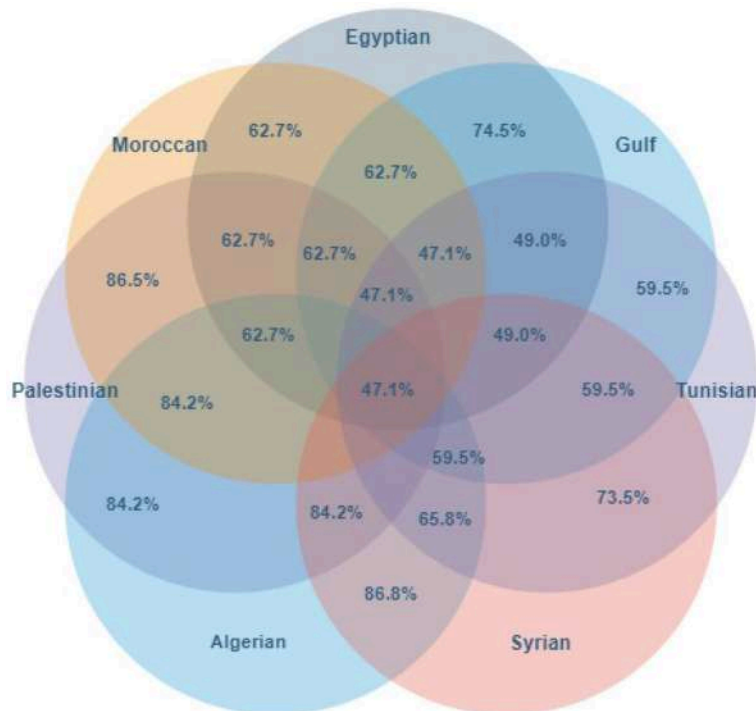
FIGURE 8
The overlap (percentage of shared lexicalizations) for Arabic dialects.

TABLE 6 The count of the diversity items collected and identified in the Indonesian languages.

Languages	Words	Gaps w/o new concepts	New concepts	Gaps considering new concepts
Indonesian	11	173	0	176
Javanese	17	167	0	170
Banjarese	12	172	3	172
Total	41	511	3	517

TABLE 7 Validator evaluation of words and lexical gaps by language.

Languages	Correctness of native speaker contribution	
	Words (%)	Gaps (%)
Indonesian	90.91	98.27
Javanese	94.44	95.78
Banjarese	91.7	97.67
Average	92.35	97.24

subdomains: siblings, cousins, and grandchildren. The total number of new concepts, 19, is lower than the sum of new concepts per language due to overlaps across languages: for example, أخٌ في الرضاعة meaning *breastfeeding brother* was found in all seven dialects, لأمّ أختٌ meaning *maternal sister* was found both in Syrian and in Egyptian, while أبيه meaning *elder cousin, son of mother's brother* only exists in Egyptian.

Validation was carried out in two phases; in the first phase, words and gaps were validated at the lexicon level by the first author, a Ph.D. student in lexical semantics and a native speaker of Arabic, and the third author, an Arabic native speaker with linguistic-semantic experience and good knowledge in Arabic dialects. In the second phase, new concepts are verified and approved to be added to the concept layer of the UKC by the second author, a lexical-semantic expert, and the UKC system manager.

Using the lexicon-level validation method, the first author evaluated the collected data in Palestinian and Syrian, while the third author validated the remaining five dialects. Results can be seen in Table 5, whereby correctness, we understand the number of words (or gaps) validated as correct divided by the total number of words (or gaps). In the case of an incorrect word, the validator either provides a correct word or indicates it as a lexical gap. For example, for the Algerian dialect, the correctness of gathered words is 85.71% and that of gaps is 98.08%. Four Algerian words were deemed incorrect: مانّي for the meaning *maternal grandmother*, لالّة for the
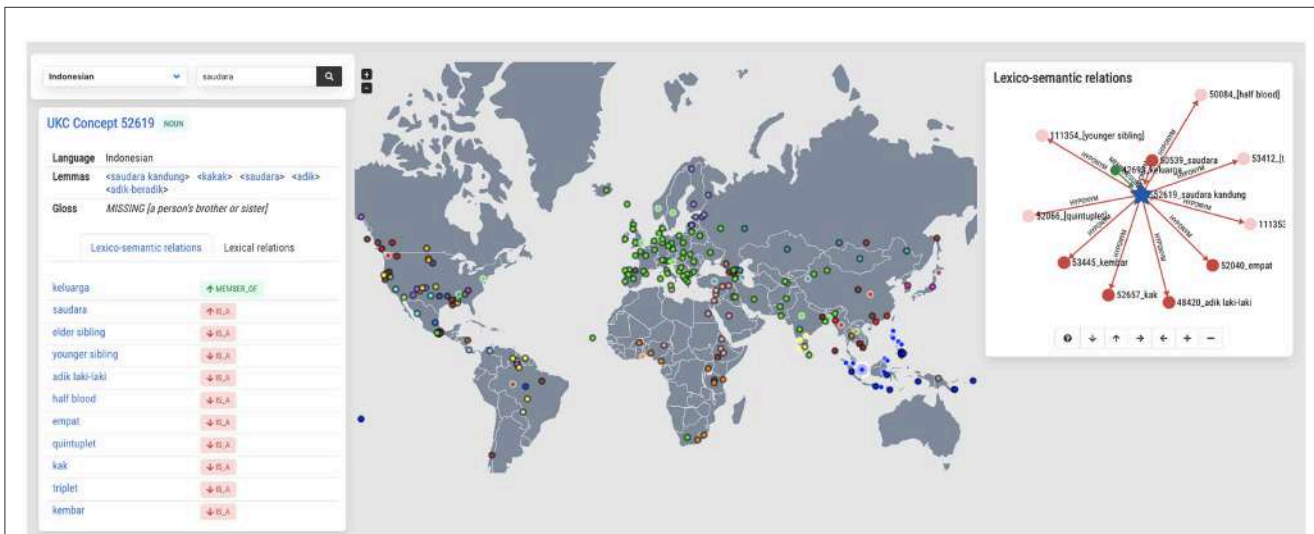
FIGURE 9
Exploring the concept of *saudara* as lexicalized in the Indonesian language **(left)**, in the world **(middle)**, and as part of the shared concept hierarchy **(right)**.
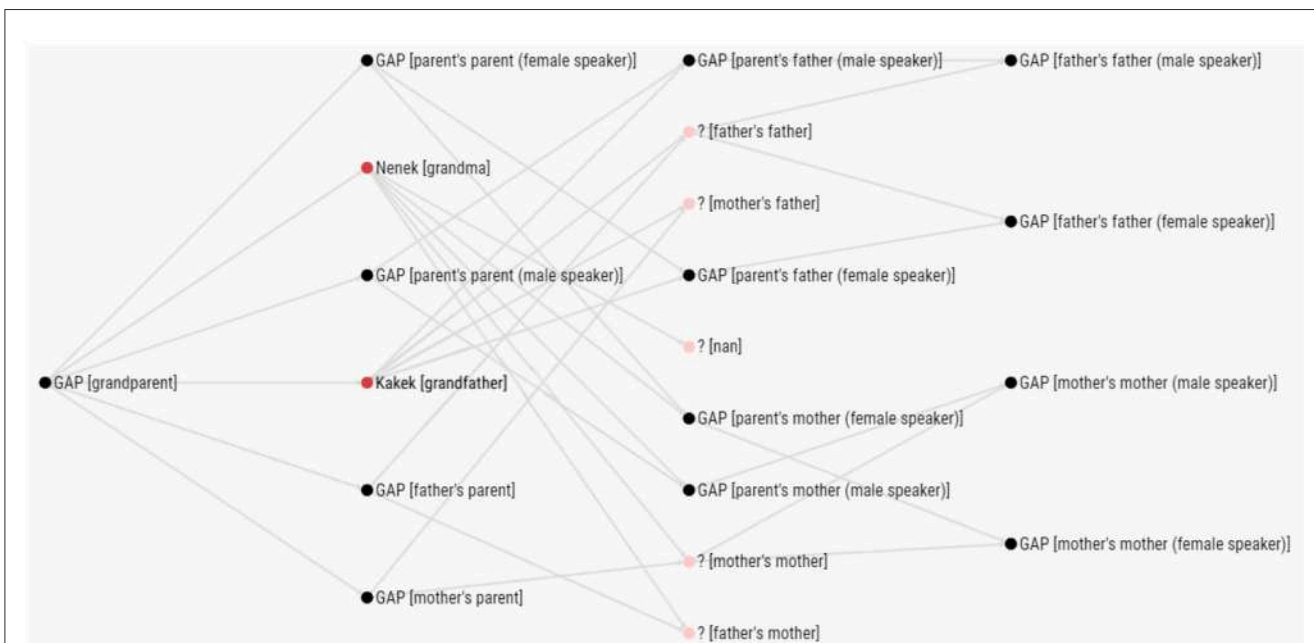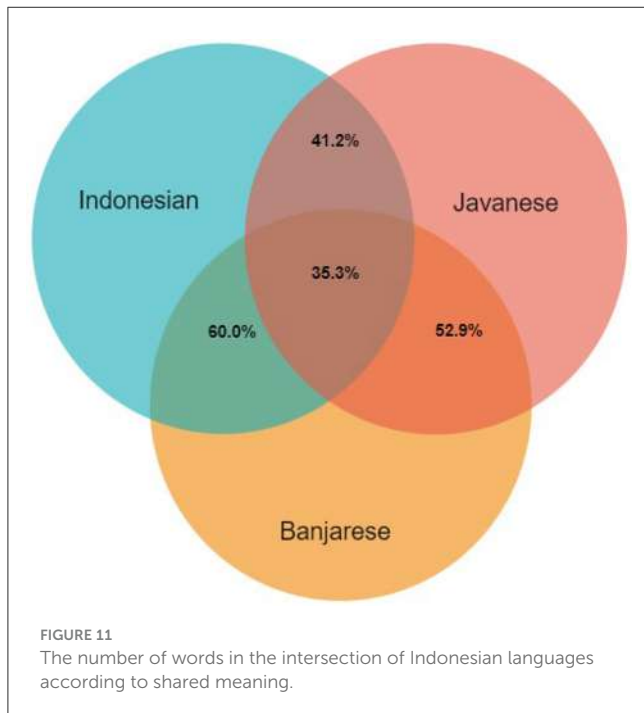

FIGURE 10
Interactive browser tool showing lexical units and gaps for the grandparent subdomain in Indonesian.

meaning *paternal grandmother*, جَدّ for the meaning *grandfather*, and باب الشيخ for the meaning *grandparent*. The validator indicated *maternal grandmother*, *paternal grandmother*, and *grandparent* as gaps, while he replaced the mistaken word جَدّ with the correct word باب الشيخ for *grandfather*. For gap evaluation, the linguistic expert validates a lexical gap by confirming it as a gap or as a non-gap due to an existing word in a dialect, for which he must provide the correct word. For instance, *Participant 1* identified the meanings *elder sister*, *father's elder sister* and *mother's elder sister* as gaps in Algerian, but the validator did not accept them and provided the polysemous word لآلّة for each of them. Evidence for validation was

obtained from the dictionary *Dictionnaire arabe algérien*[9] and from usage attested in Algerian TV films. Upon discussion between the validator and the participants, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts provided in MSA and English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

In this study, we use the UKC for creating the input dataset and the domain hierarchy and for storing and visualizing diversity

---

9 https://www.lexilogos.com/arabe_algerien.htm

FIGURE 11
The number of words in the intersection of Indonesian languages according to shared meaning.

data. Thus, the 19 new concepts were merged with the UKC by reconstructing a domain hierarchy at the supra-lingual concept layer. For example, the hierarchy of siblings was redesigned to contain five new brotherhood concepts and five new sisterhood concepts. For instance, in the Arabic-Egyptian lexicon, as shown in Figure 5, أخٌ في الرضاعة meaning "*breastfeeding brother*," is set up as a sub-node for a newly created concept of the brother, "*a male person who has the same father, mother, or both parents as another person or has the same breastfeeding woman.*", also, from the figure, can be seen أخٌ لأب meaning "*paternal brother*" and أخٌ لأمّ meaning "*maternal brother*" are inserted and connected the half-brother concept. New concepts and lexicalization are marked with white nodes and connected with blue lines.

Additionally, resulting lexical units and gaps were added into UKC lexicons. The website of the UKC provides several services for system users, such as browsable online access to database contents, source materials, and data visualization tools. The interactive exploration of linguistic diversity data in lexicons is the central feature of the website. The user can browse: (1) all meanings within a language of a word typed in by the user; and (2) lexicalizations and gaps of a concept in all languages contained in the database.

Figure 6 shows a screenshot of the concept exploration functionality describing the concept جَدّ meaning "*parent's father*". On the left-hand side of the screenshot, details are provided on the lexicalization of the concept in Arabic, such as synonymous words, a definition, and a part of speech. The middle part of the screenshot shows an interactive clickable map of all lexicons that either contain the concept or, on the contrary, lack it due to their languages being known not to lexicalize it. The color-coded dots indicate the language family, while the black circled dot represents a lexical gap. This map presents an instant global typological overview of the concept selected; for instance, from Figure 6, one can see

that most languages in Europe lexicalize the concept جَدّ while several languages in the American United States do not lexicalize it. Finally, the right-hand side shows the concept جَدّ in the context of concept hierarchy, depicted as an interactive graph: the concept, its parent and child concepts, and other lexical-semantic relations (as metonymy and meronymy) are also presented when they exist. Note that the graph only shows a part of the complete hierarchy for usability reasons. Nevertheless, it is navigable and allows the exploration of the whole concept graph in the selected language.

As mentioned at the beginning of this section, the resulting Arabic dataset will be imported into the Arabic UKC, which is an instance of the UKC system; the top layer contains independent language concepts, and the bottom layer contains twenty lexicons as the number of Arabic dialects. A screenshot of the homepage of the Arabic UKC is shown in Figure 7.

5.3 Discussion

The lexical diversity we observed across the seven dialects was higher than our original expectations, with 19 new concepts identified. Ten of these concepts are lexicalized in MSA, such as أختٌ في الرضاعة meaning "*breastfeeding sister*" and أخٌ لأب meaning "*paternal brother*". The others (nine concepts) are specific to the dialects, such as the Egyptian word أبْلَه meaning "*elder daughter of mother's sister*", which returns to the Turkish word "*kuzen*". Mostly, the origin of these Egyptian-specific concepts is the Ottoman Turkish language, when the Egyptian dialect was influenced by it during the Ottoman occupation of Egypt in the period (1517 AD to 1867 AD).

Several shared meaning overlaps have been found between dialect pairs. Likewise, intersections also existed between gaps. For a given domain d and languages $l_a, ..., l_n$, the formula below calculates the similarity of the two languages in terms of the overlap of lexicalized concepts from that domain, where LexConcepts(d, l) stands for the set of domain concepts that are lexicalized by the language l.

$$\text{overlap}(d, l_a, ..., l_n) = \frac{|\text{LexConcepts}(d, l_a) \cap ... \cap \text{LexConcepts}(d, l_n)|}{\max(|\text{LexConcepts}(d, l_a)|, ..., |\text{LexConcepts}(d, l_n)|)} \quad (1)$$

Figure 8 shows the overlaps between pairs of Arabic dialects over the kinship domain. For example, the intersection of Egyptian and Gulf dialects gives a shared coverage of 74.5%, while all dialects are 47.1% the same. In the former case, the number of lexicalizations in Egyptian is 51, and in Gulf is 42. Also, 38 of these lexical units are included in both dialects; see the dataset uploaded to GitHub.[10] For example, Formula 1 calculates the overlap between Egyptian and Gulf in the Kinship domain (K) as follows:

$$\text{overlap}(K, \text{Egyptian}, \text{Gulf}) = \frac{|\text{LexConcepts}(K, \text{Egyptian}) \cap \text{LexConcepts}(K, \text{Gulf})|}{\max(|\text{LexConcepts}(K, \text{Egyptian})|, |\text{LexConcepts}(K, \text{Gulf})|)}$$

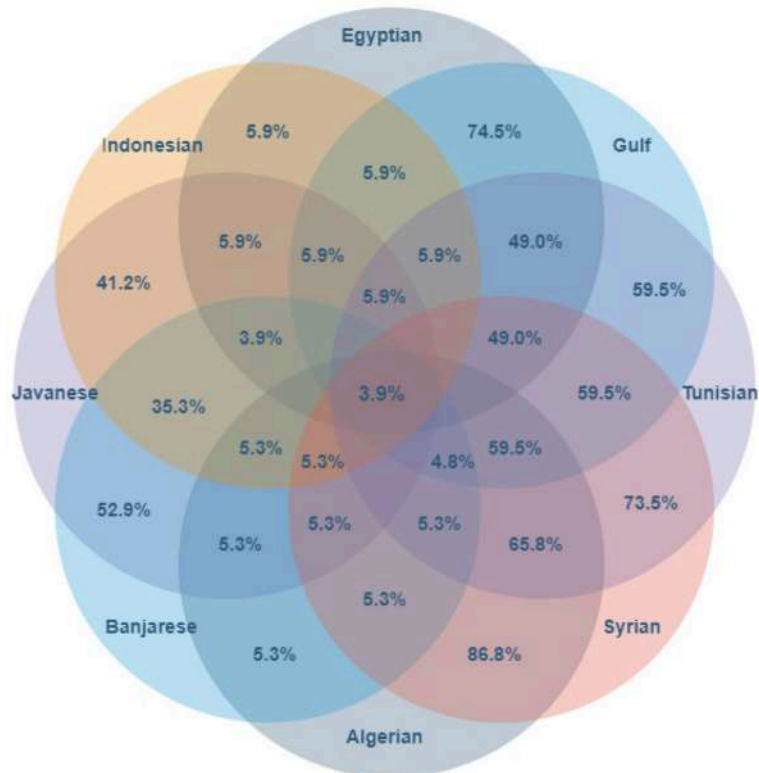10   https://github.com/HadiPTUK/kinship_dialect

FIGURE 12
The number of words in the intersection of Indonesian and Arabic languages according to shared meaning.

TABLE 8 The count of the diversity items collected and identified by domain.

Domains	Words	Gaps
Grandparents	21	169
Grandchildren	19	251
Siblings	37	173
Uncle/aunt	44	226
Nephew/niece	33	297
Cousins	67	503
Total	221	1,619

$$\text{overlap}(K, \text{Egyptian}, \text{Gulf}) = \frac{38}{\max(51, 42)} = \frac{38}{51} = 74.5\%$$

More detail about the analysis of shared coverage between the rest of the Arabic dialects can be found in the same dataset uploaded to GitHub.

We find these overlaps—e.g., an overlap of 59.5% between Gulf and Tunisian, or the overall overlap of 47.1% among all seven dialects—lower than our initial expectations on dialectal variations. Arab dialectologists justify such differences with two major factors: linguistic and religious influence (Zaidan and Callison-Burch, 2014). By linguistic influence, we refer to the historical interaction of language-speaker communities, which affects the lexicons. Examples are the Egyptian dialect influenced by the Coptic language (historically spoken by the Copts, starting from the third century

AD in Roman Egypt) or the Levantine dialect influenced by the Western Aramaic, Canaanite, Turkish, and Greek languages. The Gulf dialect is one of the Peninsular groups, which was influenced by South Arabian Languages. Secondly, the religion of the speaker community also affects the lexicon. Religion is a sociolinguistic variable that shapes how Arabic is spoken. Religion in Arab countries is a matter of group affiliation and is not usually considered an individual choice: one is born a Muslim, Christian, Jew, or Druze, and this becomes a bit like one's ethnicity. So, for example, within the Egyptian speech community, one can find language mixing between Islamic and Christian terms, and the same in the Levantine community, which consists of a mixing of Muslims, Christians, Jews, and Druze. The Gulf communities, instead, mostly consist of Muslims (Al-Wer, 2008).

6 Case study on diversity across Indonesian languages

This section demonstrates the use of the methodology described in Section 4 on kinship terminology from three Austronesian languages from Indonesia: Indonesian, Javanese, and Banjarese. Contrary to the Arabic dialects in Section 5, these three languages are not mutually intelligible.

Indonesia is the fourth most populous country in the world, and it has more than 700 living languages (Eberhard et al., 2022). The national language spoken in Indonesia is Bahasa Indonesia/Indonesian language, which was decided in the historic

moment of Youth's Pledge, October 28th, 1928. However, many Indonesians speak more than one language. For example, out of 198 million people that speak Indonesian, 84 million of them speak Javanese (Aji et al., 2022).

Even with the high number of speakers, the count of natural language processing research on Indonesian languages is very low compared to other languages around the world. As of 2020, the count of published papers on the Indonesian language is lower than other languages with less speaker count, such as Polish and Dutch (Aji et al., 2022). Not surprisingly, the amount of research on other languages (i.e., Banjarese and Javanese) in Indonesia is much lower than that. It is therefore motivating to conduct this study that discovers the richness of linguistic diversity across three Indonesian languages: standard Indonesian, Banjarese, and Javanese. In one semantic field, kinship, we have found that diversity is manifested in these languages; for example, in Javanese, the word *ponakan jaler* meaning "*nephew*", is a lexical gap in Banjarese, and in the opposite direction, the Banjarese *gulu* meaning "*parent's second eldest sibling*" is also a gap in Javanese.

6.1 Experiment setup

As in the Arabic experiment, we use the UKC lexicon to create the input dataset of kinship terms, which are independent language and formalizing such terms and also new concepts (not existing in the input dataset) identified in this experiment, as shown in the top layer of the UKC in Figure 1 for the brotherhood categorization.

In this study, three native speakers (one per language), born and educated (high school level) within the speaker community, were recruited to contribute. The participants' linguistic backgrounds are listed below:

1. *Participant 1*: a native Indonesian speaker with good command of English, Javanese, and Banjarese.
2. *Participant 2*: a native Banjarese speaker with good command of Indonesian and English.
3. *Participant 3*: a native Javanese speaker with good command of Indonesian and English.

For each language, an experiment was carried out to identify words and gaps associated with the same 184 kinship concepts as in the Arabic study (see Table 2). For example, in Banjarese, the dictionary *Kamus Bahasa Banjar Dialek Hulu-Indonesia* (Balai Bahasa Banjarmasin, 2008) and Google Search hits were used in subsequent steps to provide a precise answer on each concept from the given list of inputs. Such search steps were also followed by the Banjarese native speaker for the task of judgment on new concepts identified in the uncle/aunt subdomain. For instance, the Banjarese term *gulu*, expressing an uncle/aunt relationship with the meaning of *a parent's second eldest sibling* and attested by the dictionary above, did not previously exist in the UKC or in the KinDiv dataset, nor in Murdock (1970). Indonesian and Javanese native speakers also follow the same steps and use the dictionaries of Utomo (2015) and Badan Pengembangan dan Pembinaan Bahasa (2017) for the task of judgment on terms and gaps identified in Indonesian and Javanese, respectively.

6.2 Experiment results

The overall contribution collection effort resulted in 41 words and 517 lexical gaps. Three new, yet unattested word meanings were also found and formalized as new concepts. All three are used in Banjarese in the uncle/aunt subdomain:

- *julak*, meaning *parent's eldest sibling*;
- *gulu*, meaning *parent's second eldest sibling*;
- *angah* or *tangah*, meaning *parent's middle elder sibling* (when the number of siblings is odd).

Statistics on the data collected for each language are shown in Table 6.

As in Arabic, a two-step validation was carried out in this study. The first step validated words and gaps contributed by native speakers, carried out by the fourth author, a native Indonesian speaker with a good command of all three languages. The second validation step was done on the concept level, performed by the second author, a lexical-semantic expert and UKC system manager for new concept validation. In this step, the new concepts were verified and approved to be added to the concept layer of the UKC.

Table 7 provides correctness results over native speaker contributions, provided by the validator. Upon discussion between the validator and the contributors, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts, provided in English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

The produced kinship datasets from this experiment will be merged with the under-construction Indonesian UKC[11], a diversity-aware lexicon for languages spoken in Indonesia, also imported into the main UKC database.

Figure 9 shows how UKC explores information about a specific Indonesian word. However, the screenshot provides information about the Indonesian word *saudara*, which means "*sibling*" in English. The left-hand side of the screenshot explains synonymous words (lemmas) and the definition of the typed word. The middle of the screenshot displays the map of a global typological overview of the concept. Most languages do not lexicalize this concept, marked by the black-circled dot. Only a few languages lexicalize it, such as Indonesian, Swedish, Ainu, and Malayalam, marked by white-circled dots. The right-hand side shows the lexico-semantic relations of the concept.

The UKC lexicon is also equipped with several interactive visualization services that can be used to browse lexical units and gaps by domain in all supported languages. Figure 10 shows an example of using such services in visualizing the content of the grandparent subdomain in Indonesian.

6.3 Discussion

More than 90% of our 184 initial kinship concepts were found to be gaps in the three Indonesian languages, as shown in

---

11   http://indonesia.ukc.datascientia.eu/

Table 6. Using Formula 1, we calculated the overlaps between the Indonesian languages in terms of kinship lexicalizations, shown in Figure 11. For more details, see the dataset uploaded to the GitHub repository[12]. 35.3% of the concepts are lexicalized by the three Indonesian languages studied. The Javanese–Banjarese overlap is 52.9%, Javanese–Indonesian is 60%, and finally Banjarese–Indonesian is 41.2%. Even though all three languages are included in the Malayo-Polynesian branch of the Austronesian language family, Indonesian and Banjarese are considered Malayic languages, while Javanese is not, which is the first reason for this result. Furthermore, these languages exist on different islands in Indonesia; Javanese exists on Java Island, Banjarese is located on the southern part of Borneo Island, and the Indonesian language is based on Malay, which is spoken on Sumatra Island (Sneddon, 2003), so this geographical barrier restricts interactions between speakers, and each language has developed within its own speech community.

Finally, using Formula 1, we computed the overlaps between Arabic dialects and Indonesian languages. Figure 12 shows that the ten languages together cover only 3.9% of the concepts, and the most similar language pair, namely Egyptian–Indonesian, is merely 5.9% similar. For researchers in ethnography or comparative linguistics, the observation of such pronounced levels of cross-lingual and cross-cultural diversity may not come as a surprise, as major variations in kin patterns are well-known in these domains. On the other hand, we believe that beyond these narrow fields of research, there is a general lack in understanding the depth of diversity in how, through languages, people describe and interpret the world. Most computational linguists and engineers who build language processing systems, as well as the users who trust such systems for their daily activities, do not suspect the breadth of the mental divide across languages that language applications, such as machine translation systems, are meant to bridge. We think that through quantified measures, as we are attempting to do with our simple measure of overlap introduced on p. 18, can be useful to improve our qualitative grasp on diversity, which we consider a promising direction for future research.

Table 8 includes statistics of collected words and gaps by domain across Arabic and Indonesian languages. The results show that only three words in the domain of cousins are identified in the Indonesian languages, while in Egyptian, 16 words are used around the concept of the cousin.

7 Related work

Ethnologists and linguists have for a long time studied how family structures map to kinship terminology across languages and social groups. The most famous and comprehensive ethnographic study on kin term patterns is that of Murdock (1970), upon which our work also indirectly relied: our cross-lingual formalization of kin terms is based on the one provided by the KinDiv resource, itself in part derived from Murdock's data. KinDiv covers 699 languages and is a computer-processable database that can also be exploited for applications in computational linguistics. Our results provide

linguistic evidence in seven Arabic dialects and three Indonesian languages that do not figure in these resources.

The exploration of kin terminology and the building of large-scale databases on the topic has also been the subject of more recent efforts—we only cite two examples here. The AustKin project[13] has produced a large-scale database on kin terms in hundreds of indigenous Australian languages. The recent Kinbank database (Passmore et al., 2023) is a comprehensive resource on kinship terminology, covering over 1,173 languages, with a broad coverage of kinship subdomains. As Kinbank was released after the initial submission of our paper, we did not rely on it for our work. We consider our research as complementary to Kinbank: concentrating on a relatively low number of dialects and languages, our results could, in principle, be integrated into Kinbank in order to extend its coverage. And vice versa, we see potential in using Kinbank data in order to cross-validate and possibly to extend the Indonesian terms we collected (as the three Indonesian languages of our study are also covered by Kinbank). There is, however, an important methodological difference between the our and Kinbank's way of representing terms: Kinbank does not explicitly indicate lexical gaps. For example, our work considers the concept of *son of father's brother as pronounced by a male speaker* to be a lexical gap in Javanese, while Kinbank maps the Javanese term *sedulur misan*, simply meaning *cousin*, to this and 95 other meanings. Our work, instead, identifies the Javanese term as the general meaning of *cousin* and considers all other (more specific) cousin terms as lexical gaps. This distinction is useful in comparative linguistics and cross-lingual applications where the explicit indication of the lack of precise meaning equivalence can be exploited.

Concepticon (List et al., 2016) is 'a resource for linking concept lists' frequently used in comparative linguistics. The *concept sets* of Concepticon serve the same purpose as the supra-lingual concepts of the UKC in our study, namely to provide meaning-based mappings among lists of terms (aka *concept lists* in Concepticon) across languages. As of mid-2023, Concepticon consists of nearly 4,000 concept sets, principally targeting core vocabularies (basic-level categories) that are the main subject of study of historical and comparative linguistics. Concepticon is under continuous development and has more recently evolved from a flat list of meanings to a hierarchy with broader–narrower relations. At the time of writing, the kinship domain seems to be partially represented in Concepticon: while sibling or grandparent relations are widely covered, fine-grained cousin relationships are mostly missing from it. The UKC, which contains over 100,000 supra-lingual concepts and a wide range of lexical and lexico-semantic relations, was a more suitable resource for our study due to its more complete coverage of the kinship domain and its explicit support for representing term untranslatability via lexical gaps.

Multilingual computational applications being in the core of our focus, we also review relevant resources from computational linguistics. For NLP applications, the most popular and widely-known representation of lexico-semantic knowledge is that of *wordnets* that follow the general structure of the original English *Princeton WordNet* (Miller, 1995). The *wordnet expansion* approach by Fellbaum and Vossen (2012)—an expert-driven lexicon

---

12   https://github.com/HadiPTUK/kinship_dialect

13   http://austkin.net

translation effort—is frequently used to produce new wordnets for lower-resourced languages: this approach consists of 'translating' (i.e., finding lexicalizations for) English WordNet concepts ('synsets' in wordnet terminology) into the target language. While this is a straightforward approach that produces resources that remain cross-lingually linked, its downside is that the translation approach cannot involve concepts and words specific to the target language and not present in the source language (which in most cases is English). In cases of diverse conceptualizations of the world, the translation approach often results in incorrect approximations. To take the example of Arabic, both versions of the Arabic Wordnet (Elkateb et al., 2006; Abouenour et al., 2013) map the English synset of *uncle* ("*the brother of your father or mother; the husband of your aunt*") to the Arabic synset of عَمّ, which means "*the brother of your father.*"

A similar situation is observed for Indonesian. As far as we know, the only Indonesian Wordnet currently accessible is Bahasa Wordnet—a bilingual Wordnet for standard Indonesian and Malay languages (Noor et al., 2011). It was formed by merging three different wordnets (one in Indonesian and two in Malay) developed mainly by the same expansion approach from PWN. Due to this approach, many English words that have no equivalents in Indonesian are incorrectly mapped, resulting in meaning loss. For example, in Bahasa Wordnet, the English word *sister*, which means "*a female person who has the same parents as another person*," was mapped to the Indonesian word *kakak* which means "*elder sibling.*"

Finally, we mention MultiWordNet as an early effort at improving the representation of linguistic diversity in multilingual lexical databases (Pianta et al., 2002). It is a multilingual lexicon that was built using the *merge* method that, contrary to the translation-based expand approach presented above, maps together existing high-quality bilingual dictionaries. MultiWordNet explicitly represents lexical gaps in its Italian and Hebrew wordnets: about 1,000 in Italian and about 300 in Hebrew (Bentivogli and Pianta, 2000; Ordan and Wintner, 2007). MultiWordNet, however, is a discontinued effort that does not cover the kinship domain and is thus was not suitable for our purposes.

The methodology we present in Section 4 follows neither the expansion nor the merge approach but a third one, more adapted to diversity-aware lexicography: our starting point is a supra-lingual, diversity-aware conceptualization of the domain of study (kinship in our case). The task of *contribution collection* is performed by native speakers with respect to the supra-lingual concept hierarchy based on evidence from comparative linguistics and covering a wide range of languages. While there is no guarantee that our initial conceptualization is complete—indeed, it was not the case in our study—it is less biased toward the concepts of a single language and speaker community than the expansion approach.

8 Conclusions and future work

Our paper formally captures lexical diversity across languages and dialects by representing language- or dialect-specific concepts and linguistic gaps. It introduces a systematic method to produce such data in a human-based manner from one semantic domain rather than from general domains, as the efforts of covering the

WordNet domains (Magnini and Cavaglià, 2000) that have been conducted in building these wordnets, Mongolian (Batsuren et al., 2019), Unified Scottish Gaelic (Bella et al., 2020), and MultiWordNet (Pianta et al., 2002).

The method is verified through two large-scale case studies on kinship terminology, a domain known to be diverse across languages and cultures: one case study deals with seven Arabic dialects, while the other one with three Indonesian languages. The experiments show that our method outperforms the existing methods in terms of the quantity of explored gaps and words and the quality of results. Overall efforts resulted in 1619 gaps, and 223 words were identified in 10 languages and dialects. Moreover, 22 new word meanings with respect to the imported list of independent-language concepts from the UKC are explored in this research.

In future work, we plan to automate the method presented in this paper and apply it to new languages, such as the rest of the Arabic dialects and Indonesian language, as well as to new domains that are known to be diverse, such as body parts, food, color, or visual objects (Giunchiglia and Bagchi, 2021; Giunchiglia et al., 2023).

Finally, diversity-aware lexicons such as the UKC (which includes our produced datasets) provide essential information to cross-lingual applications, such as multilingual NLP tasks or cross-lingual language models. In the future, we plan to use this resource in implementing one such application, i.e., machine translation.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://github.com/HadiPTUK/kinship_dialect.

Author contributions

FG and GB conceptualized and supervised the study. GB and HK imported and formatted the dataset of inputs. HK wrote the original manuscript draft and performed the Arabic experiments. AF and HK validated the collected Arabic data at the lexicon level. SD performed the Indonesian experiments and validated the results at the lexicon level. GB validated the identified diverse data at the concept level. FG, GB, AF, and HK analyzed the Arabic and Indonesian data. FG, GB, AF, SD, and HK reviewed and edited the manuscript. All authors contributed to the research and approved the submitted version.

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1229697/full#supplementary-material

References

Abouenour, L., Bouzoubaa, K., and Rosso, P. (2013). On the evaluation and improvement of Arabic WordNet coverage and usability. *Lang. Resour. Eval.* 47, 891–917. doi: 10.1007/s10579-013-9237-0

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., et al. (2022). "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin: Association for Computational Linguistics), 7226–7249.

Al-Wer, E. (2008). "Arabic languages, variation in," in *Concise Encyclopedia of Languages of the World*, eds K. Brown and S. Ogilvie (Oxford: Elsevier Ltd.), 53–56.

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., et al. (2018). "A cross-linguistic database of phonetic transcription systems," in *Yearbook of the Poznan Linguistic Meeting* (Poznań: De Gruyter Open), 21–53.

Arora, A., Farris, A., Gopalakrishnan, R., and Basu, S. (2021). "Bhāṣācitra visualising the dialect geography of South Asia," in *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021* (Association for Computational Linguistics), 51–57.

Badan Pengembangan dan Pembinaan Bahasa (2017). *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan.

Balai Bahasa Banjarmasin (2008). *Kamus Bahasa Banjar Dialek Hulu-Indonesia*. Banjarbaru: Departemen Pendidikan Nasional, Pusat Bahasa, Balai Bahasa Banjarmasin.

Batsuren, K., Bella, G., and Giunchiglia, F. (2019). "CogNet: a large-scale cognate database," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 3136–3145.

Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., et al. (2022). "UniMorph 4.0: universal morphology," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 840–855.

Bella, G., Batsuren, K., Khishigsuren, T., and Giunchiglia, F. (2022a). "Linguistic diversity and bias in online dictionaries," in *Frontiers in African Digital Research*, ed K. Lena (Bayreuth: Institute of African Studies), 173–186.

Bella, G., Byambadorj, E., Chandrashekar, Y., Batsuren, K., Cheema, D., and Giunchiglia, F. (2022b). "Language diversity: visible to humans, exploitable by machines," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Dublin: Association for Computational Linguistics), 156–165.

Bella, G., Helm, P., Koch, G., and Giunchiglia, F. (2023). Towards bridging the digital language divide. *arXiv preprint arXiv:2307.13405*. doi: 10.48550/arXiv.2307.13405

Bella, G., McNeill, F., Gorman, R., Donnaíle, C. Ó., MacDonald, K., Chandrashekar, Y., et al. (2020). "A major Wordnet for a minority language: Scottish Gaelic," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 2812–2818.

Bentivogli, L. and Pianta, E. (2000). "Looking for lexical gaps," in *Proceedings of the 9th EURALEX International Congress*, eds U. Heid and S. Evert (Stuttgart: Institut fur Maschinelle Sprachverarbeitung), 663–669.

Carling, G., Larsson, F., Cathcart, C. A., Johansson, N., Holmer, A., Round, E., et al. (2018). Diachronic Atlas of Comparative Linguistics (DiACL)–a database for ancient language typology. *PLoS ONE* 13, e0205313. doi: 10.1371/journal.pone.0205313

Catford, J. C. (1965). *A Linguistic Theory of Translation*. London: Oxford University Press.

Dryer, M. S., and Haspelmath, M. (eds.). (2013). *WALS Online (v2020.3)*. Zenodo.

Eberhard, D., Simons, G. F., and Fenning, C. D. (2022). *Ethnologue: Languages of Africa and Europe*. Dallas, TX: SIL International Publications.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., et al. (2006). "Building a WordNet for Arabic," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (Genoa: European Language Resources Association), 29–34.

Fellbaum, C., and Vossen, P. (2012). Challenges for a multilingual WordNet. *Lang. Resour. Eval.* 46, 313–326. doi: 10.1007/s10579-012-9186-z

Georgakopoulos, T., Grossman, E., Nikolaev, D., and Polis, S. (2022). Universal and macro-areal patterns in the lexicon: a case-study in the perception-cognition domain. *Linguist. Typol.* 26, 439–487. doi: 10.1515/lingty-2021-2088

Giunchiglia, F., Bagchi, M., and Diao, X. (2023). A semantics-driven methodology for high-quality image annotation. *arXiv preprint arXiv:2307.14119*.

Giunchiglia, F., and Bagchi, M. (2021). Classifying concepts via visual properties. *arXiv preprint arXiv:2105.09422*. doi: 10.48550/arXiv.2105.09422

Giunchiglia, F., Batsuren, K., and Bella, G. (2017). "Understanding and exploiting language diversity," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (Melbourne, VIC), 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, CiCling2018*, ed A. Gelbukh (Hanoi: Springer), 18–24.

Helm, P., Bella, G., Koch, G., and Giunchiglia, F. (2023). Diversity and language technology: how techno-linguistic bias can cause epistemic injustice. *arXiv preprint arXiv:2307.13714*. doi: 10.48550/arXiv.2307.13714

Kay, P., and Cook, R. S. (2016). "World color survey," in *Encyclopedia of Color Science and Technology*, eds M. R. Luo (New York, NY: Springer), 1265–1271.

Kemp, C., and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science* 336, 1049–1054. doi: 10.1126/science.1218811

Khishigsuren, T., Bella, G., Batsuren, K., Freihat, A. A., Chandran Nair, N., Ganbold, A., et al. (2022). "Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 2798–2807.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., et al. (2016). D-PLACE: a global database of cultural, linguistic and environmental diversity. *PLoS ONE* 11, e0158391. doi: 10.1371/journal.pone.0158391

Kopecka, A., and Narasimhan, B. (2012). *Events of Putting and Taking: A Crosslinguistic Perspective*. Amsterdam: John Benjamins Publishing.

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

Levinson, S. C., and Wilkins, D. P. (2006). *Grammars of Space: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

List, J.-M., Cysouw, M., and Forkel, R. (2016). "Concepticon: a resource for the linking of concept lists," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož: European Language Resources Association), 2393–2400.

Magnini, B., and Cavaglià, G. (2000). "Integrating subject field codes into WordNet," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)* (Athens: European Language Resources Association).

Majid, A., Bowerman, M., van Staden, M., and Boster, J. S. (2007). The semantic categories of cutting and breaking events: a crosslinguistic perspective. *Cogn. Linguist.* 18, 133–152. doi: 10.1515/COG.2007.005

McCarthy, A. D., Wu, W., Mueller, A., Watson, B., 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) (Hong Kong: Association for Computational Linguistics), 2241–2250.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM* 38, 39–41.

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

Muttaqin, Z. (2009). Fiqh lughah dalam literatur Arab klasik. *Afaq 'Arabiyah: Jurnal Kebahasaaraban dan Pendidikan Bahasa Arab* 4, 107–122.

Noor, N. H. B. M., Sapuan, S., and Bond, F. (2011). "Creating the open Wordnet Bahasa," in *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation* (Tokyo: Institute of Digital Enhancement of Cognitive Processing, Waseda University), 255–264.

Ordan, N., and Wintner, S. (2007). Hebrew WordNet: a test case of aligning lexical databases across languages. *Int. J. Transl.* 19, 39–58.

Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., Sheard, C., Argyriou, P., et al. (2023). Kinbank: a global database of kinship terminology. *PLoS ONE* 18, e0283218. doi: 10.1371/journal.pone.0283218

Pianta, E., Bentivogli, L., and Girardi, C. (2002). "Developing an aligned multilingual database," in *Proceedings of the 1st International WordNet Conference* (Mysuru: Global Wordnet Association), 293–302.

Plungyan, V. (2011). Modern linguistic typology. *Herald Russian Acad. Sci.* 81, 101–113. doi: 10.1134/S1019331611020158

Reznikova, T., Rakhilina, E., and Bonch-Osmolovskaya, A. (2012). Towards a typology of pain predicates. *Linguistics* 50, 421–465. doi: 10.1515/ling-2012-0015

Roberson, D., Davidoff, J., Davies, I. R., and Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cogn. Psychol.* 50, 378–411. doi: 10.1016/j.cogpsych.2004.10.001

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., et al. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data* 7, 1–13. doi: 10.1038/s41597-019-0341-x

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

Sneddon, J. (2003). *The Indonesian Language*. Sydney, NSW: University of New South Wales Press Ltd.

Utomo, S. S. (2015). *Kamus Indonesia-Jawa*. Jakarta: PT Gramedia Pustaka Utama.

Viberg, Å. (1983). The verbs of perception: a typological study. *Linguistics* 21, 123–162.

Wälchli, B., and Cysouw, M. (2012). Lexical typology through similarity semantics: toward a semantic map of motion verbs. *Linguistics* 50, 671–710. doi: 10.1515/ling-2012-0021

Wierzbicka, A. (2007). Bodies and their parts: an NSM approach to semantic typology. *Lang. Sci.* 29, 14–65. doi: 10.1016/j.langsci.2006.07.002

Zaidan, O. F., and Callison-Burch, C. (2014). Arabic dialect identification. *Comput. Linguist.* 40, 171–202. doi: 10.1162/COLI_a_00169